Análisis de Técnicas de Detección de Imágenes Sintéticas

Analysis of Synthetic Image Detection Techniques

TRABAJO FIN DE MÁSTER
MÁSTER EN INGENIERÍA INFORMÁTICA
CURSO 2023–2024

Daniel Cabañas González

Director: **Luis Javier García Villalba**
Colaboradora: **Ana Lucila Sandoval Orozco**

Departamento de Ingeniería del Software e Inteligencia Artificial
Facultad de Informática
Universidad Complutense de Madrid

Madrid, Junio de 2024

Convocatoria: junio-julio
Nota final: 10 (Matrícula de Honor)

A mis padres y a mis gatas,

"Aprender a apreciar lo que la vida nos ofrece
nos permite descubrir en qué consiste la felicidad."

– Martha E. González

Agradecimientos

En primer lugar, quiero expresar mi sincero agradecimiento a mis directores y al grupo de investigación GASS cuyos componentes han facilitado significativamente el progreso de esta investigación.

Quiero extender un profundo agradecimiento a mi familia, cuyo apoyo y ánimos han sido fundamentales a lo largo de mi trayectoria académica. Gracias por confiar en mí, por quererme y por motivarme cuando lo he necesitado.

Quisiera expresar mi más profunda gratitud a Sara, por su constante apoyo, sus consejos y su capacidad para adaptarse a mis extenuantes horarios. Gracias por estar a mi lado, por tu paciencia y cariño.

Quisiera hacer una mención especial a Gonzalo, a Ignacio, a Pablo y a Guillermo, mis mejores amigos a los que considero mis hermanos.

Un agradecimiento especial a Guillermo Pérez Tamayo por su mentoría y apoyo continuo durante los últimos años. Tu sabiduría y guía han sido esenciales en mi crecimiento académico y profesional.

Finalmente, deseo expresar mi gratitud a todos los compañeros, profesores y amigos que me han acompañado a lo largo del máster. Quiero hacer una mención especial a Santi, Carlos, Jesús, Álex y Ovi. A cada uno de vosotros y al resto, mi más sincero agradecimiento por vuestro apoyo y compañía durante este periodo.

Índice General

Índice de Figuras

Índice de Tablas

Lista de Acrónimos

AR	Auto-Regressive Model
CelebA	Celebrity Faces Attributes Dataset
CNN	Convolutional Neural Network
DCT	Discrete Cosine Transform
DDIM	Denoising Diffusion Implicit Model
DDPM	Denoising Diffusion Probabilistic Model
DFDC	Deepfake Detection Challenge
DFF	DeepFakeFace
DFT	Discrete Fourier Transform
DIRE	Diffusion Reconstruction Error
DM	Diffusion Model
EBM	Energy-Based Model
FFHQ	Flickr Faces HQ
FFT	Fast Fourier Transform
GAN	Generative Adversarial Network

HP-FFT	High Pass - Fast Fourier Transform
HP-DFT	High Pass - Discrete Fourier Transform
ICCV	International Conference on Computer Vision
LAION	Large-scale Artificial Intelligence Open Network
LDM	Latent Diffusion Model
LSUN	Large-scale Scene UNderstanding
MNIST	Modified National Institute of Standards and Technology database
MSCOCO	Microsoft Common Objects in Context
ResNet	Residual Neural Network
SVM	Support Vector Machine
VAE	Variational AutoEncoder

Abstract

This research addresses the study and analysis of the current state of detectors for synthetic images generated by diffusion models. Through a methodology that includes the study of previous generative models and the application of advanced detection techniques such as binary classifiers and frequency spectrum analysis, this work explores the efficacy of current methods and their applicability to diffusion models. The results reveal that, although the classifiers are effective within the datasets with which they were trained, they face significant limitations in generalizing to new models, highlighting the need for more robust and generalizable tools. The conclusions emphasize the urgency of continuing research in this field, given the serious implications of deepfakes and unregulated synthetic content. This study not only validates the possibility of adapting existing techniques to new generative models but also provides a starting point for the development of practical solutions in the fight against cybercrime, specifically in the detection and prevention of illegal and harmful content. The tools developed from this research could be essential for security forces and other entities dedicated to combating the pernicious effects of deepfake technology.

Keywords: AI, datasets, deepfakes, diffusion models, GANs, synthetic images detection.

Resumen

Esta investigación aborda el estudio y análisis del estado actual de los detectores de imágenes sintéticas generadas por modelos de difusión. A través de una metodología que incluye el estudio de modelos generativos previos y la aplicación de técnicas de detección avanzadas como clasificadores binarios y análisis de espectro de frecuencia, este trabajo explora la eficacia de los métodos actuales y su aplicabilidad a los modelos de difusión. Los resultados revelan que, aunque los clasificadores son efectivos dentro de los conjuntos de datos con los que fueron entrenados, enfrentan limitaciones significativas al generalizar a nuevos modelos, evidenciando la necesidad de herramientas más robustas y generalizables. Las conclusiones destacan la urgencia de continuar con investigaciones en este campo, dada la seriedad de las implicaciones de los deepfakes y contenido sintético no regulado. Este estudio no solo valida la posibilidad de adaptar técnicas existentes a nuevos modelos generativos, sino que también ofrece un punto de partida para el desarrollo de soluciones prácticas en la lucha contra el cibercrimen, específicamente en la detección y prevención de contenido ilegal y dañino. Las herramientas desarrolladas a partir de esta investigación podrían ser esenciales para cuerpos de seguridad y otras entidades dedicadas a combatir los efectos perniciosos de la tecnología deepfake.

Palabras Clave: IA, conjuntos de datos, deepfakes, modelos de difusión, GANs, detección de imágenes sintéticas.

Capítulo 1

Introducción

1.1. Motivación

Durante las últimas décadas, hemos sido testigos de una revolución digital que ha transformado drásticamente el mundo tecnológico, particularmente en el ámbito de la creación y manipulación de imágenes. El desarrollo de modelos generativos como las redes GAN, los VAE y, más recientemente, los modelos de difusión, ha marcado un avance significativo en cómo podemos generar imágenes sintéticas de alta calidad que son casi indistinguibles de las reales.

En los primeros días de la inteligencia artificial aplicada a la generación de imágenes, las limitaciones de hardware y software restringían la calidad y la complejidad de las imágenes generadas. Sin embargo, con la evolución de la tecnología digital, estos modelos han mejorado en gran medida, resultando en imágenes que no solo engañan al ojo humano, sino que también presentan desafíos significativos para su detección mediante métodos convencionales.

El salto hacia modelos más avanzados como los modelos de difusión ha complicado aún más la tarea de diferenciar entre imágenes reales y sintéticas. Aunque inicialmente se diseñaron para mejorar la calidad de las imágenes generadas, su capacidad para producir resultados hiperrealistas ha abierto nuevas puertas para aplicaciones tanto positivas como potencialmente maliciosas, como la creación de desinformación o contenido falso. La capacidad de los modelos de difusión para generar imágenes indistinguibles de las reales ha llevado a un aumento alarmante en la producción de contenido inapropiado, incluyendo material pornográfico que involucra representaciones de menores. Este tipo de contenido, aunque sintético, es profundamente problemático y sigue siendo ilegal en muchas jurisdicciones del mundo. La facilidad y accesibilidad con que estos modelos pueden generar imágenes hiperrealistas han resultado en un incremento exponencial de material de abuso sexual infantil en el ámbito digital, lo cual es profundamente alarmante y requiere una acción urgente.

Por otro lado, aunque los métodos de detección desarrollados para modelos generativos anteriores han demostrado ser relativamente eficaces, adaptándose bien a las características específicas de las imágenes que generan, estas técnicas encuentran limitaciones significativas cuando se aplican a los modelos de difusión, que emplean mecanismos diferentes y más sofisticados para la creación de imágenes. Dado este escenario, se hace imprescindible no solo adaptar los métodos de detección existentes, sino también

desarrollar nuevas técnicas que puedan identificar de manera efectiva las complejidades de las imágenes generadas por los últimos modelos.

Este trabajo se propone estudiar en profundidad la situación actual de la generación de imágenes sintéticas, especialmente aquellas producidas por modelos de difusión, que representan el frente más avanzado en tecnologías generativas. Asimismo, se evaluará el estado actual de las técnicas de detección de contenido sintético para identificar sus limitaciones y áreas de mejora. A partir de esta evaluación, el objetivo es desarrollar, probar y proponer soluciones que mejoren la capacidad de identificar y discriminar entre imágenes reales y sintéticas generadas por los modelos de difusión. La finalidad de este esfuerzo es doble: por un lado, mejorar la seguridad digital y, por otro, proporcionar herramientas confiables para el análisis forense de imágenes en un contexto dominado por la inteligencia artificial.

1.2. Contexto

El presente Trabajo Fin de Máster se enmarca dentro de un proyecto de investigación titulado Novel Strategies to Fight Child Sexual Exploitation and Human Trafficking Crimes and Protect their Victims – HEROES, aprobado por la Comisión Europea dentro del Programa Marco Horizonte 2020 (convocatoria H2020-SU-SEC-2020) en virtud del acuerdo de subvención número 101021801 y en el que participa como coordinador del proyecto el Grupo GASS de la Universidad Complutense de Madrid (Grupo de Análisis, Seguridad y Sistemas, https://gass.ucm.es, grupo 910623 del catálogo de grupos de investigación reconocidos por la UCM).

Además de la Universidad Complutense de Madrid participan en HEROES 24 entidades ubicadas en 17 países: 11 de países de la UE (Austria, Bélgica, Bulgaria, Francia, Grecia, Irlanda, Letonia, Lituania, Portugal, España, Reino Unido), 1 país asociado (Suiza) y 5 terceros países (Bangladesh, Brasil, Colombia, Perú, Uruguay). Dichas entidades son: University of Kent (Reino Unido), The Free University of Brussels (Bélgica), The French National Research Institute for Digital Science and Technology – INRIA (Francia), Center for Security Studies – KEMEA (Grecia), International Centre for Migration Policy Development – ICMPD (Austria), International Center for Missing and Exploited Children – ICMEC (Suiza), IDENER Research & Development Agrupación de Interés Económico (España), Athena Research Center – ARC (Grecia), Trilateral Research and Consulting (Reino Unido), Centre for Women and Children Studies – CWCS (Bangladesh), Center Against Human Trafficking and Exploitation – KOPZI (Lituania), Portuguese Association for Victim Support – APAV (Portugal), Fundación Renacer (Colombia), The Greek Council for Refugees – GCR (Grecia), Brazilian Association for the Defense of Children of Children and Youth – ASBRAD (Brasil), Hellenic Police (Grecia), Latvia National Police (Letonia), General Directorate for the Fight against Organized Crime (Bulgaria), Dirección General de la Policía – DGP (España), Federal Police (Brasil), Federal Highway Police (Brasil), Secretaría de Inteligencia Estratégica de Estado – Presidencia de la República Oriental del Uruguay (Uruguay)

1.3. Objeto de la Investigación

La detección de imágenes sintéticas generadas mediante modelos de difusión plantea desafíos únicos en comparación con modelos generativos anteriores como las GANs. Aunque los modelos de difusión producen imágenes de alta calidad y realismo, investigaciones recientes han demostrado que, al igual que sus predecesores, también dejan trazas y artefactos detectables que pueden ser explotados para diferenciar entre imágenes reales y sintéticas. Estas trazas pueden no ser perceptibles a simple vista, pero son identificables a través de técnicas avanzadas de análisis de imágenes.

Este trabajo se centra en el estudio de estas trazas residuales específicas que dejan los modelos de difusión, explorando cómo las redes neuronales convolucionales con arquitecturas como ResNet pueden ser entrenadas para clasificar imágenes en términos de su autenticidad. Estas redes son capaces de aprender y reconocer patrones complejos en datos visuales que a menudo son invisibles para los humanos, ofreciendo una herramienta poderosa en la lucha contra la desinformación visual y los deepfakes.

Además, se abordará el análisis de imágenes en el dominio de Fourier, una técnica que permite estudiar las frecuencias de las imágenes para detectar variaciones sutiles que no son evidentes en el espacio temporal. Este enfoque es particularmente útil para identificar las firmas específicas que los modelos de difusión dejan en las imágenes sintéticas, como anomalías en las texturas o patrones inusuales en los objetos contenidos en las imágenes, los cuales son indicativos de contenido sintético.

La combinación de estos enfoques metodológicos tiene como objetivo desarrollar un sistema de detección robusto que pueda ser utilizado no solo para identificar imágenes generadas por modelos de difusión, sino también para proporcionar una base sólida para el desarrollo de futuras herramientas de detección a medida que evolucionan estas tecnologías. Este estudio contribuirá significativamente al campo de la seguridad digital, proporcionando conocimientos y recursos para combatir efectivamente el uso malintencionado de imágenes sintéticas.

1.4. Plan de Trabajo

El desarrollo de este trabajo se plantea en cuatro fases:

1. **Investigación**: La fase de investigación se centrará en los primeros cuatro meses, donde el proyecto se dedicará a comprender y aprender el contexto de los modelos generativos. Durante este período, se llevará a cabo un estudio profundo de la literatura existente sobre modelos generativos como GANs, VAEs y Transformers, y un análisis específico y extenso de los modelos de difusión. Se realizarán reuniones periódicas para discutir los avances, resolver dudas y compartir recursos útiles como Google Scholar para la búsqueda de artículos relevantes. Esta fase también incluirá la exploración de datos y técnicas previas en la detección de imágenes sintéticas, con el objetivo de destacar la escasez de información específica sobre modelos de difusión, lo que delineará el camino hacia la experimentación y desarrollo de nuevas herramientas.

2. **Desarrollo**: Tras consolidar la base teórica, se transitará hacia la fase de desarrollo, con el objetivo de centrar esfuerzos en la generación de un conjunto de datos

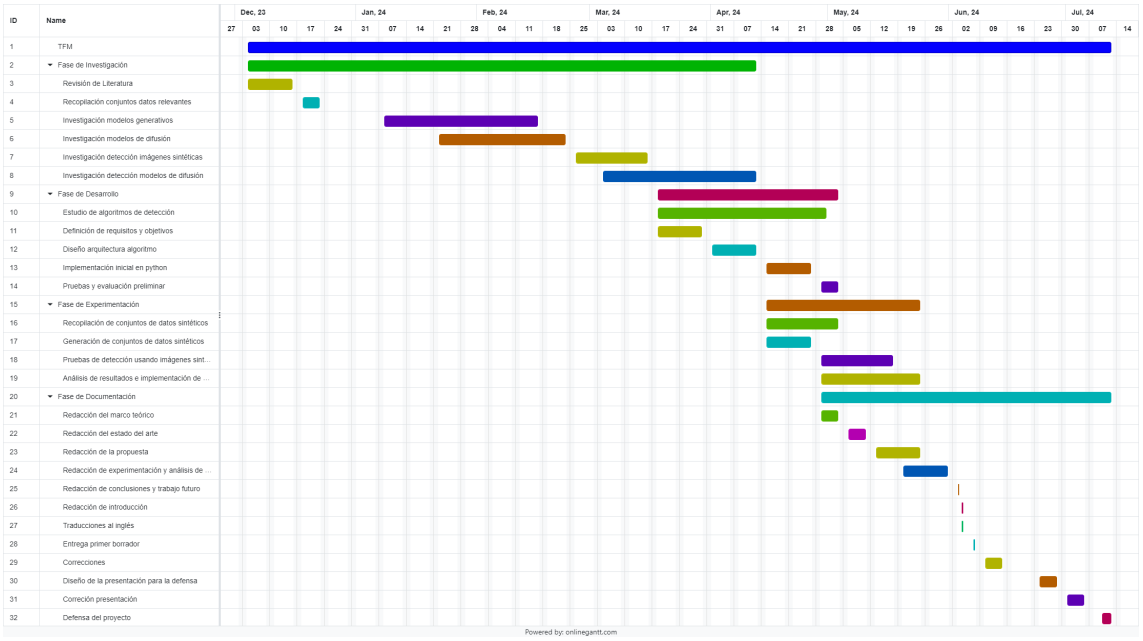


Figura 1.1: Diagrama de Gantt del proyecto, desde su concepción hasta la defensa final.

propio usando herramientas de difusión latente recientes como Stable Diffusion [RBL+22], dado el limitado acceso a datasets específicos de imágenes generadas por modelos de difusión. Además, se prepararán y ajustarán arquitecturas de redes neuronales como ResNet para entrenar con estos nuevos conjuntos de datos. Durante esta etapa, se aplicarán transformaciones en el dominio de Fourier para observar detalladamente las trazas y artefactos residuales generados por los modelos de difusión, técnicas utilizadas para identificar características distintivas en las imágenes sintéticas producidas por modelos generativos anteriores.

3. **Experimentación**: La tercera fase involucrará la implementación práctica de clasificadores binarios para probar su eficacia en la discriminación de imágenes reales contra sintéticas. Se experimentará con diferentes configuraciones de ResNet, ajustando parámetros y evaluando su rendimiento con múltiples datasets para garantizar que los modelos puedan generalizar y detectar más de un solo tipo de imagen sintética. Se realizarán comparaciones de los resultados obtenidos y se reajustarán en las arquitecturas y parámetros de los modelos con el objetivo de optimizar estos modelos.

4. **Documentación**: En la última fase se dedicará tiempo a documentar cada etapa del proceso, desde la fase de investigación inicial hasta las pruebas finales de los clasificadores, con el objetivo de crear un documento de memoria completo para el Trabajo de Fin de Máster, sirviendo como una colección de todo el trabajo realizado y como un recurso valioso para futuras investigaciones en el campo de la detección de imágenes sintéticas.

La Figura 1.1 presenta un diagrama de Gantt que desglosa las fases del proyecto y sus respectivas duraciones ofreciendo una representación gráfica clara de la distribución del tiempo a lo largo del proyecto.

1.5. Estructura del Trabajo

El resto del trabajo está organizado en seis capítulos y dos apéndices con la estructura que se comenta a continuación: El Capítulo 2 introduce algunos conceptos que son elementales para comprender los modelos generativos, explorando distintos tipos como las GANs, CNNs, VAEs, ARs, Transformers, y en particular, los modelos de difusión. Se discuten los procesos de difusión y reversión y se describen los diferentes tipos de modelos de difusión que existen, estableciendo la base teórica necesaria para los siguientes capítulos.

El Capítulo 3 aborda el estado actual de la detección de imágenes sintéticas. El capítulo comienza con una revisión de cómo funcionan las técnicas de detección en modelos generativos anteriores y luego se centra en la detección de imágenes generadas por modelos de difusión. Se discute la aplicabilidad de técnicas existentes para modelos anteriores y la adaptación necesaria para los modelos de difusión. También se analizan los conjuntos de datos disponibles, destacando la escasez de datos para modelos de difusión y el análisis de imágenes en el dominio de Fourier.

El Capítulo 4 presenta la metodología empleada en este proyecto, describiendo los clasificadores binarios, las arquitecturas ResNet, y el uso de distintas transformaciones para el análisis de frecuencia. Se expone cómo se prepararon los conjuntos de datos utilizados, incluyendo la generación de un conjunto propio, y la planificación del proyecto para abordar los objetivos de investigación establecidos.

El Capítulo 5 describe los experimentos realizados para evaluar la efectividad de los algoritmos propuestos en el capítulo 4. Este capítulo cubre desde la configuración de los entrenamientos de modelos usando diversas arquitecturas ResNet y técnicas de aumento de datos, hasta la evaluación de estos modelos usando distintos conjuntos de datos. Se incluye un análisis detallado de los resultados, comparando diferentes configuraciones y también se aplican análisis de frecuencia para comprobar las trazas que dejan los modelos de difusión en las imágenes sintéticas.

El Capítulo 6 se presentan las conclusiones derivadas de la investigación, resumiendo los hallazgos principales y discutiendo las implicaciones de estos. También se esbozan las líneas futuras de investigación, proponiendo cómo estos estudios pueden expandirse y adaptarse para seguir mejorando la detección de imágenes sintéticas en un campo en constante evolución.

Los Capítulos 7 y 8 son las traducciones al inglés de la Introducción y de las Conclusiones.

El Apéndice A muestra el funcionamiento de un script propuesto para generar un conjunto de datos de imágenes sintéticas generadas por Stable Diffusion [RBL+22].

El Apéndice B recolecta resultados generados en esta investigación con el objetivo de proporcionar imágenes de mayor calidad y tamaño para poder apreciar de manera correcta trazas de artefactos que producen los modelos de difusión en el dominio de Fourier.

Capítulo 2

Fundamentación Teórica

En este capítulo se introducirá al lector los conceptos básicos sobre los modelos generativos, los cuales son modelos de aprendizaje automático diseñados para aprender distribuciones de datos y generar nuevas instancias estadísticamente similares a los datos originales. Entre los datos que pueden producir los modelos generativos se encuentran las imágenes sintéticas, una práctica que ha evolucionado de manera notable en los últimos años.

En el desarrollo de este capítulo, se examinarán en detalle diversos tipos de modelos generativos, abarcando desde las redes generativas antagónicas hasta los modelos de difusión. Inicialmente, en la Sección 2.1, se introducirán los fundamentos de los modelos generativos, discutiendo su evolución y las distintas aplicaciones. Más adelante, en la Sección 2.2 se profundizará en los modelos de difusión, para acabar hablado de los tipos de modelos de difusión existentes en la Sección 2.3.

2.1. Introducción a los modelos generativos

Durante los últimos años se han desarrollado múltiples metodologías para la generación de información sintética utilizando técnicas de inteligencia artificial para ello. Entre estas técnicas se encuentran las llamadas Redes Generativas Antagónicas, que hasta hace muy poco eran el método preferido para la generación de imágenes sintéticas debido a su rapidez y su eficiencia. También se pueden encontrar el uso de modelos de difusión y sus variantes, las Redes Neuronales Convolucionales, Autoencoders Variacionales, Modelos Autoregresivos, Transformes y Modelos Basados en Energía entre otros:

- **GANs**: Las Redes Generativas Antagónicas (Generative Adversarial Network (GAN)) son redes de aprendizaje profundo que se utilizan para la generación de información artificial y en su mayoría para la generación de imágenes sintéticas. Una red GAN está compuesta por dos redes neuronales que entrenan de manera simultánea en un sistema basado en la competencia, donde la primera red (conocida como el generador) se dedica a generar datos sintéticos lo más parecido a los datos reales que recibe como entrada, y la segunda red (conocida como el discriminador) debe distinguir entre los datos reales y los datos generados por la primera red.

  Este modelo fue propuesto inicialmente en [GPAM+14] y desde entonces ha sido utilizado para una amplia gama de aplicaciones, desde la generación de imágenes

sintéticas y su mejora en la calidad y resolución [KALL17], la generación de ejemplos de entrenamientos para otros sistemas de inteligencia artificial [Ba19].

Las GANs han sido hasta hace poco la herramienta más utilizada para la generación de imágenes sintéticas y "deepfakes" [KM18], imágenes, videos y audios donde las caras, voces y otros atributos de personas reales son reemplazados por contenido sintético. Los deepfakes se popularizaron inicialmente con un carácter humorístico, pero poco después se empezaron a utilizar con fines maliciosos, como pueden ser la manipulación de información y la suplantación de identidad [KM19, FDSP+19].

- **CNNs**: Las Redes Neuronales Convolucionales (Convolutional Neural Network (CNN)) son un tipo de redes neuronales profundas que normalmente se utilizan para la clasificación y reconocimiento de imágenes, pero también pueden ser utilizadas para generar imágenes sintéticas, como se puede ver en el desarrollo del PixelCNN [VdOKE+16]. El uso de CNNs para la generación de imágenes sintéticas conlleva una serie de limitaciones, como la necesidad de reajustes y post-procesamiento, que evitaron su popularización para este tipo de tareas.

- **VAEs**: Los Autoencoders Variacionales (Variational AutoEncoder (VAE)) son un tipo de modelo generativo basado en redes neuronales que han sido diseñadas para aprender representaciones de información comprimida y luego reconstruir datos a partir de estas representaciones [PGH+16]. Un VAE está compuesto por dos partes diferenciadas. La primera de ellas consiste en un codificador o encoder, que se encarga de recibir los datos de entrada y transformarlos en una distribución de espacio latente (en una distribución probabilística comúnmente de tipo normal) para poder trabajar con la variabilidad e incertidumbre de los datos de entrada. La segunda parte se conoce como decodificador o decoder, y es la encargada de tomar muestras de la distribución generada por el encoder y a partir de estos intentar reconstruir los datos originales. Los VAEs son utilizados para múltiples aplicaciones, desde la generación de imágenes, sonidos y texto hasta el refuerzo de otros sistemas semi-supervisados o no supervisados [ZZCH18].

- **ARs**: Un Modelo Autorregresivo (Auto-Regressive Model (AR)) es un tipo de modelo estadístico que se basa en la descripción de procesos temporales en series de tiempo, donde los valores de los datos se obtienen de una combinación lineal de valores anteriores más un término de error. En el contexto de los modelos generativos, los ARs generan imágenes de manera secuencial prediciendo el valor de los píxeles a partir de los píxeles de las secuencias anteriores [YXK+22].

- **Transformers**: Un Transformer en el contexto de modelos generativos consiste en una arquitectura específica de red neuronal avanzada, utilizada ampliamente en el procesamiento del lenguaje natural y en la generación de imágenes y audio. Su principal característica es el uso de lo que se conoce como mecanismo de atención [VSP+17], que les permite centrarse en partes específicas de los datos de entrada para poder calcular los datos de salida. También pueden implementar un modelo de codificación y decodificación como hacen los VAEs. Los Transformers también han servido proponiendo una arquitectura de red que posteriormente ha sido adaptada en modelos futuros [DYH+21, CWG+21].

- **EBMs**: Los Modelos Basados en Energía (Energy-Based Model (EBM)) son modelos de aprendizaje profundo que contienen una función de energía la cual asigna un valor

escalar a cada dato de entrada, cuya intención es reflejar cómo de probable es una determinada configuración (cuanta menos energía, más favorable) [LCH+06].

- **DMs**: Los Modelos de Difusión (Diffusion Model (DM)) son una clase de modelos generativos probabilísticos basados en dos etapas. La primera de las etapas consiste en la difusión de información de manera gradual, añadiendo ruido a los datos en cada paso hasta lograr que la información sea completamente ruido. La segunda etapa consiste en la reversión, eliminando el ruido hasta recuperar los datos originales [SDWMG15, HJA20]. Los modelos de difusión han adquirido gran relevancia en los últimos años debido a varios factores como la calidad con la que generan imágenes, las amplias posibilidades de configuración durante la generación y la flexibilidad que estos modelos ofrecen [CHIS23, YZS+23].

2.2. Modelos de Difusión

En el contexto de los modelos generativos, la aparición de los modelos de difusión supone un nuevo paradigma a la hora de abordar la generación de datos sintéticos a partir de datos reales, ya que estos modelos son capaces de generar imágenes con una alta calidad y poseen una gran robustez en sus procesos de entrenamiento [HJA20]. Este tipo de modelos empieza a destacar frente a los modelos más célebres del momento, las Redes Generativas Antagónicas (GANs), que aun siendo capaces de generar imágenes de alta calidad encuentran limitaciones en sus fases de entrenamiento como son los colapsos o las salidas limitadas [SC21]. Los Autoencoders Variacionales (VAEs) en cambio, ofrecen estabilidad en el entrenamiento a costa de producir imágenes que no poseen tanta calidad [PGH+16] como la que ofrecen los modelos de difusión o las Redes Generativas Antagónicas.

2.2.1. ¿Qué son los Modelos de Difusión?

Como se ha introducido anteriormente, los modelos de difusión son un tipo de modelo generativo probabilístico que utilizan los procesos de difusión y reversión para crear nuevos datos. Los procesos de difusión y reversión emulan la manera que tienen de propagarse y deshacerse ciertos fenómenos físicos como son la difusión de partículas en un medio [SDWMG15].

A diferencia de otros tipos de modelos generativos, como las Redes Generativas Antagónicas (GANs) o los Autoencoders Variacionales (VAEs), los modelos de difusión no utilizan una función de mapeo de espacio latente a espacio de datos, sino que se dedican a la degradación de la información existente mediante el proceso de difusión para luego revertir este proceso con el objetivo de generar datos nuevos a partir de ruido aleatorio [HJA20, Luo22]. Este proceso tiene como objetivo lograr un balance entre la calidad de la imagen producida y la estabilidad durante el entrenamiento, con el objetivo de lograr resultados de alta calidad sin llegar a toparse con las dificultades que se encuentran en los otros modelos generativos, y además proporciona una capacidad de configuración más explícita en el proceso generativo y permitiendo ajustes finos difíciles de lograr con los otros modelos [YZS+23].

2.2.2. Procesos de difusión y reversión

Uno de los primeros avances en el contexto de los modelos de difusión y también relevante para el desarrollo del aprendizaje profundo no supervisado tuvo lugar tras la publicación del artículo [SDWMG15] en 2015, donde se explicó que los principios de la termodinámica, y en particular aquella fuera del equilibrio, podría aplicarse al aprendizaje automático. En el artículo los autores hablan sobre la simulación de dinámicas de fluidos y sobre difusión de partículas bajo un marco probabilístico. En la Figura 2.1 se pueden observar los procesos de difusión y reversión detallados en el artículo [SDWMG15].

Alejándose de los modelos tradicionales de aprendizaje automático, los autores del artículo proponen un proceso de aprendizaje como una serie de transformaciones aplicadas a los datos de una manera parecida a cómo la energía se disipa en los sistemas físicos. El foco se estableció en la observación de cómo los datos evolucionan en función del tiempo bajo una dinámica de difusión probabilística, normalmente gaussiana, que introduciría ruido en los datos de manera gradual hasta alcanzar un estado de ruido puro, comprendiendo este proceso como un proceso reversible.

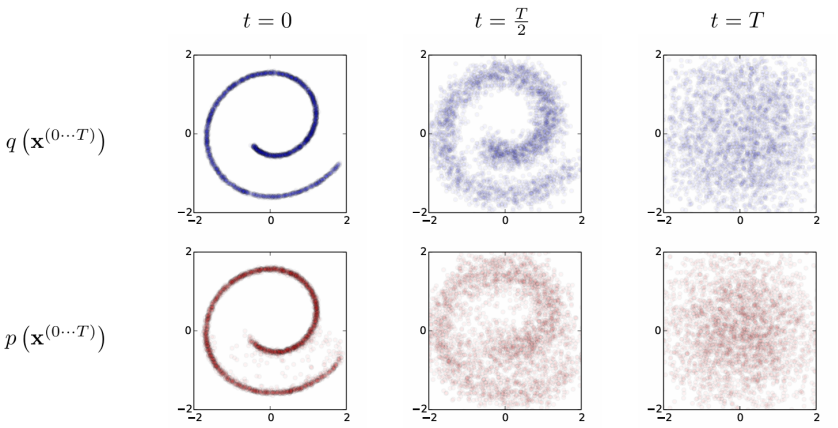


Figura 2.1: Procesos de difusión (en azul) y reversión (en rojo) propuestos en [SDWMG15] aplicados a un conjunto de datos con forma de rollo suizo (swiss roll).

Para poder comprender mejor el proceso de difusión, uno se puede imaginar qué pasaría si deja caer una gota de colorante en un vaso lleno de agua y espera un cierto tiempo. Al rato, el colorante se habría diluido y esparcido por el agua de manera gradual. Se comprende este proceso de difusión con la capacidad de ser revertido y que, tras una serie de transformaciones y aplicaciones que se verán más adelante, se podría lograr deshacer la dilución para poder "reconstruir" la gota de colorante. El imaginar este proceso que consiste en deshacer la dilución ayuda a entender mejor el proceso de reversión.

En [SDWMG15], los autores ejemplifican un proceso de difusión gaussiano (en azul) añadiendo ruido progresivamente hasta difuminar de manera completa e irreconocible un conjunto de datos en forma de espiral como se muestra en la Figura 2.1. En ella también se puede observar el proceso de reversión (en rojo) basado en un proceso de difusión gaussiano invertido donde previamente se conocían las funciones de media y desviación típica.

Según [SDWMG15], el proceso de difusión es descrito como una cadena de Markov para transformar gradualmente una distribución gaussiana hasta convertirla en una distribución compuesta por puro ruido. Cada paso de esta cadena de Markov tiene una probabilidad evaluable de manera analítica, y por lo tanto, la cadena entera también es analíticamente

evaluable. El proceso consiste en la aplicación de un kernel de difusión de Markov, $T\pi$, que difunde los datos utilizando una tasa de difusión β, como se puede ver en la Ecuación (2.1):

$$q(x^{(t)}|x^{(t-1)}) = \mathcal{N}(x^{(t)}; x^{(t-1)}\sqrt{1-\beta_t}, \beta_t \mathbf{I}) \tag{2.1}$$

Ecuación 2.1: Cadena de Markov para la aplicación de ruido gaussiano descrita en los anexos de [SDWMG15].

En esta ecuación, $x^{(t)}$ representa los datos en el paso de tiempo t, β_t es un parámetro que determina la cantidad de ruido gaussiano añadido en cada paso, e $\mathbf{I}$ es la matriz de identidad.

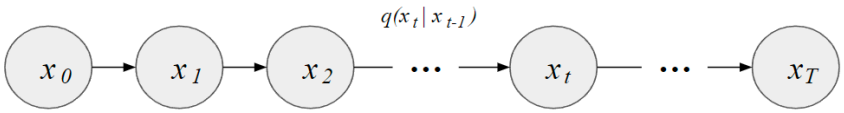
El proceso de reversión se entrena para poder recuperar la estructura inicial. El modelo generativo basa su trayectoria de manera reversa y usa transiciones Markovianas aprendidas con media y covarianza conocidas, como se puede ver en la Ecuación (2.2):

$$p(x^{(t-1)}|x^{(t)}) = \mathcal{N}(x^{(t-1)}; \mu_\theta(x^{(t)}, t), \Sigma_\theta(x^{(t)}, t)) \tag{2.2}$$

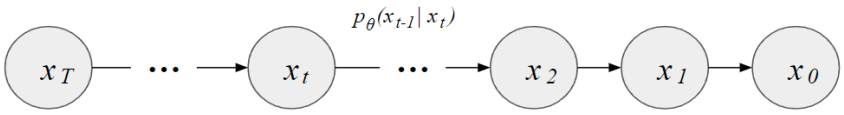
Ecuación 2.2: Cadena de Markov para la reversión del ruido gaussiano previamente aplicado, descrita en los anexos de [SDWMG15].

Se puede observar que en la Ecuación (2.2) μ_θ representa la media y Σ_θ la covarianza, las cuales serían aprendidas por una red neuronal en el entrenamiento.

El artículo [SDWMG15] estableció un marco para la modelización probabilística mediante la aplicación de procesos de difusión y reversión y se establecieron las bases de los modelos generativos basados en modelos de difusión.



(a) Proceso de difusión.



(b) Proceso de reversión.

Figura 2.2: Diagrama que representa la cadena de Markov del proceso de difusión y reversión

En 2020 se publicó el artículo [HJA20], donde se utilizaron los conceptos establecidos en [SDWMG15] para presentar el primer modelo de generación de imágenes sintéticas. En el artículo [HJA20], los autores implementan la aplicación de procesos de difusión a

imágenes, incrementando gradualmente el ruido en éstas en una cadena de Markov hasta difuminar completamente la imagen con ruido (ver Figura 2.2(a)), para luego aplicarles el proceso de reversión, lo que ellos denominan "desruido", con el objetivo de poder recuperar la imagen inicial (ver Figura 2.2(b)). Un ejemplo práctico de la aplicación del proceso de difusión a una imagen se puede encontrar en la Figura 2.3.

Los autores proponen la posibilidad de entrenar modelos profundos usando una cadena de Markov parametrizada con inferencia variacional para poder producir muestras que coincidan con los datos en un tiempo finito. Las transacciones de esta cadena de Markov son aprendidas para poder revertir el proceso de difusión.

Los autores definen la forma en la que se introduce el ruido gradualmente en los datos originales utilizando la Ecuación (2.3):

$$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}), \quad q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{2.3}$$

Ecuación 2.3: Proceso de difusión descrito en [HJA20].

Esta fórmula define un proceso secuencial donde cada paso t se construye a partir del paso anterior $t-1$, transformando los datos originales x_0 en una serie de datos cada vez más ruidosos hasta llegar al último paso T.

- $x_{1:T}$ representa una secuencia de estados de datos desde el tiempo 1 hasta el tiempo T.
- x_0 es estado inicial de la secuencia antes de aplicar la difusión.
- $q(x_{1:T} \mid x_0)$ es la distribución de probabilidad de la secuencia de estados $x_{1:T}$ dada la muestra inicial x_0.
- $q(x_t \mid x_{t-1})$ es la distribución de probabilidad de transición de un estado x_{t-1} a x_t.
- $\mathcal{N}(x_t; \sqrt{1-\beta_t})$ es una distribución normal (gaussiana) que define cómo se añade ruido a los datos en cada paso de tiempo.
- β_t es un parámetro que controla la cantidad de ruido añadida en cada paso y va aumentando con el tiempo, llevando a los datos a ser más ruidosos.
- $\beta_t I$ representa la varianza de la distribución normal, donde I es la matriz identidad, asegurando que el ruido añadido tenga la misma varianza en todas las dimensiones de los datos.

Los autores definen la forma en la que el modelo aprende a revertir el proceso de difusión, utilizando transiciones gaussianas aprendidas por el modelo profundo para reconstruir los datos originales como se puede ver en la Ecuación (2.4):

$$p_\theta(x_{0:T}) := p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t), \quad p_\theta(x_{t-1} \mid x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{2.4}$$

Ecuación 2.4: Proceso de reversión del ruido gaussiano descrito en [HJA20].

De manera similar a la Ecuación (2.3), esta fórmula define el proceso secuencial donde cada paso t se deduce a partir del paso anterior $t-1$, aplicando el proceso de reversión desde x_T hasta eliminar todo el ruido aplicado y poder recuperar los datos originales x_0.

- $p_\theta(x_{0:T})$ es la distribución conjunta sobre toda la secuencia de estados $x_{0:T}$ bajo el modelo generativo θ.
- $p(x_T)$ es la distribución de probabilidad para el estado final x_T.
- $\prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t)$ es el producto de todas las distribuciones de transición $p_\theta(x_{t-1} \mid x_t)$ de $t = 1$ a T.
- $p_\theta(x_{t-1} \mid x_t)$ representa la distribución de probabilidad de un estado x_{t-1} dado el siguiente estado x_t bajo el modelo generativo θ.
- $\mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$ es una distribución normal con media $\mu_\theta(x_t, t)$ y covarianza $\Sigma_\theta(x_t, t)$, que modela la transición de x_t a x_{t-1}.

La Ecuación (2.4) representa el modelo generativo $p_\theta(x_{0:T})$ que busca aproximar la distribución de datos reales $q_\theta(x_{0:T})$ minimizando la divergencia Kullback-Leibler (KL) entre el modelo p_θ y la distribución real q.

El artículo [HJA20] supuso un avance significativo en la generación y reconstrucción de datos, y en el contexto de los modelos generativos debido a la potencia y eficiencia lograda por los entonces nuevos modelos de difusión, siendo estos una alternativa más estable y eficiente que las técnicas anteriores como las GANs o las VAEs, estableciendo un nuevo estándar en la generación de imágenes sintéticas.

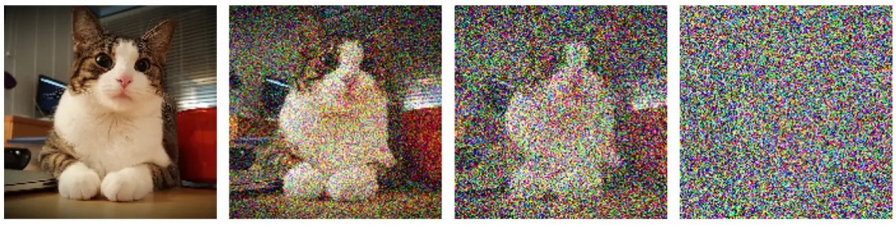


Figura 2.3: Ejemplo de proceso de difusión mediante la aplicación de ruido gaussiano de manera progresiva [VK22].

2.3. Tipos de Modelos de Difusión

El primer modelo de difusión propuesto fue el mencionado anteriormente en el artículo [HJA20], conocido como Modelo Probabilístico de Difusión con Desruido (Denoising Diffusion Probabilistic Model (DDPM)). Durante los últimos años se han propuesto distintas variantes de modelos de difusión que introducen distintas técnicas tanto en el entrenamiento de las redes neuronales como en los procesos de difusión [CHIS23], y que se verá a continuación.

2.3.1. Denoising Diffusion Probabilistic Models

Estos modelos fueron los primeros en aparecer, como se ha visto en el artículo [HJA20]. En estos modelos se simula el proceso de difusión visto en el apartado anterior, añadiendo ruido de manera gradual a los datos de entrada hasta convertirlos completamente en ruido. Las redes neuronales entrenan para aprender a revertir el ruido en cada paso, generando nuevas instancias de datos que tienen como objetivo imitar de la manera más similar a los datos originales. Se puede encontrar un diagrama que ejemplifica este proceso en la Figura 2.5.

Cabe destacar que en los DDPM implementan el uso de lo que se conoce como "Redes U" (U-net), arquitecturas de redes neuronales profundas, cuyo esquema tiene forma de U (ver Figura 2.4), utilizadas debido a su capacidad para capturar y procesar información en diferentes niveles de resolución. El concepto de Redes U fue implementado por primera vez en 2015 en el artículo [RFB15]. Estás Redes U serán utilizadas como columna vertebral dentro de muchos modelos generativos, sobre todo aquellos orientados a la generación de imágenes sintéticas.

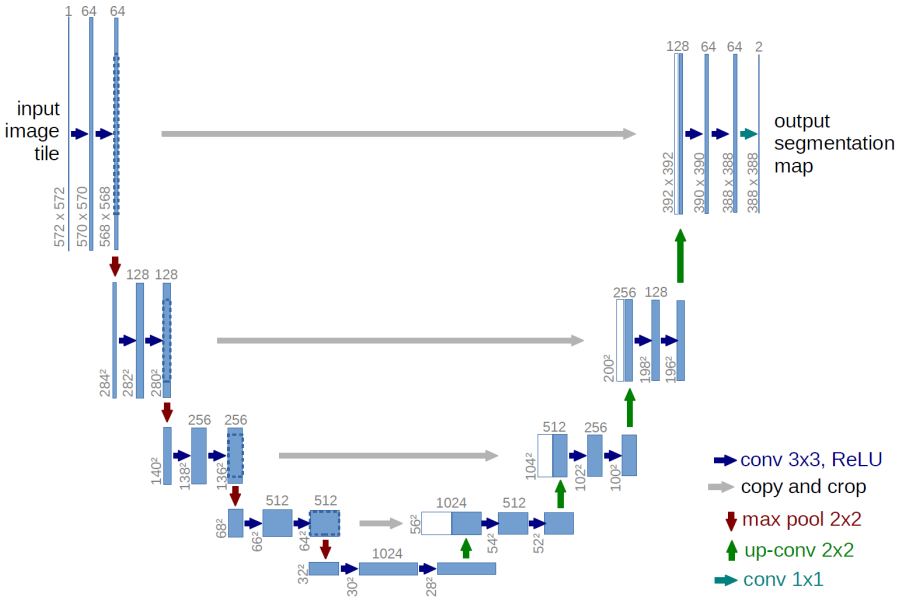


Figura 2.4: Diagrama de una Red U propuesta en [RFB15]

Como ya se ha comentado en la sección anterior, se demostró en varios artículos como [HJA20, ND21] que las DDPM eran capaces de producir imágenes y otro tipos de datos sintéticos con una muy alta calidad y con un nivel de detalle impresionante, superando en muchos casos a otros modelos generativos como las GANs, las VAEs y las ARs.

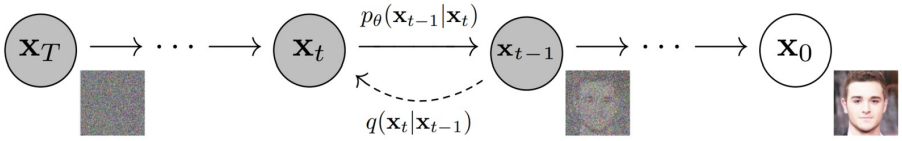


Figura 2.5: Proceso de desruido del primer DDPM representado gráficamente en el artículo [HJA20].

En las DDPM también se implementó la técnica de "Inpainting", una técnica de

edición y restauración de imágenes donde se reconstruyen partes de las imágenes o vídeos. Estas partes pueden haber estado dañadas, distorsionadas o simplemente recortadas con el objetivo de generar nuevas variantes. La técnica de "Inpainting" es utilizada frecuentemente en el contexto de los modelos generativos, pues es interesante poder modificar parcialmente una imagen en vez de tener que generarla desde cero. Para ello, se necesita delimitar las partes que van a ser generadas y existen distintos enfoques y herramientas para ello, desde el uso de máscaras hasta la delimitación de franjas mediante la indicación de los píxeles y otros parámetros [YLY+18].

Esta técnica es una de las técnicas más utilizadas para la generación de perfiles sintéticos, ya que permite generar rostros nuevos de manera sencilla en imágenes ya existentes sin tener que generar las propias imágenes. En los artículos [ZWS+22, NKH19] se comenta la técnica de "Inpainting" utilizada por redes GANs y el artículo [LDR+22] es uno de los artículos más conocidos acerca la técnica de "Inpainting" en los DDPM.

2.3.2. Denoising Diffusion Implicit Models

Los Modelos Implícitos de Difusión con Desruido (Denoising Diffusion Implicit Model (DDIM)) [SME20] son una variante de los DDPM que ofrecen un enfoque alternativo y más eficiente que sus predecesores para simular y revertir el proceso de difusión. Los DDIM implementan un modelo determinista en el proceso de reversión, a diferencia de los DDPM que utilizan un modelo estocástico [HJA20], por lo que en los procesos de reversión de los DDIM estos requieren un menor número de pasos, siendo estos más eficaces e informativos. Los DDIM también permiten un control más directo sobre la trayectoria de muestreo, siendo esta más optimizable mediante el uso de parámetros.

Con la aparición de las DDIM también se introdujeron antiguas técnicas como la superresolución [GLZ+23] o el Inpaiting [ZJZ+23], y se presentaron variantes propias de los DDIM como se comenta en [ZTC22].

2.3.3. Latent Diffusion Models

Los Modelos de Difusión Latentes (Latent Diffusion Model (LDM)) son una variante de los modelos de difusión que aparecieron en 2022 a raíz de la publicación del artículo [RBL+22].

Los LDMs trabajan utilizando un espacio comprimido respecto al espacio original, que es conocido como "espacio latente". El espacio latente es una representación de menor dimensionalidad que los datos originales, y esta técnica se logra mediante el uso de autoencoders. El objetivo de trabajar en el espacio latente es el de trabajar con datos significativamente menores, los cuales se procesan más rápidamente y con menor coste computacional. Los LDMs aprenderían a generar nuevas instancias de datos e imágenes en este espacio comprimido que más adelante se decodificarían para lograr recuperar el espacio de los datos e imágenes originales, y con esto se lograría generar imágenes de alta calidad. Los autores del artículo introducen así el concepto de difusión latente.

En el artículo [RBL+22] los autores explican los beneficios de utilizar el enfoque de la difusión latente, que consisten en la generación de imágenes con muy alto nivel de detalle y complejidad con un coste de procesado muy eficiente, un rendimiento competitivo, la posibilidad de implementar la super-resolución en imágenes y vídeos, la generación de imágenes desde texto y la reducción de costes computacionales respecto a los modelos de

difusión tradicionales.

Dos técnicas a destacar en la implementación de este nuevo enfoque de difusión latente serían el uso de autoencoders para la generación de los espacios latentes y el uso de capas de atención cruzadas que permiten el manejo de entradas condicionales como son los textos o los esquemas (ver Figura 2.6). Los LDMs adquirieron gran relevancia sobre todo en el ámbito de la generación de imágenes para diagnósticos médicos [PTD+22, TN23] debido a su capacidad para generar imágenes de muy alta calidad, aunque también destacaron en la generación de vídeos sintéticos [BRL+23] y la generación de datos sintéticos a partir de entradas de texto [RLJ+23].

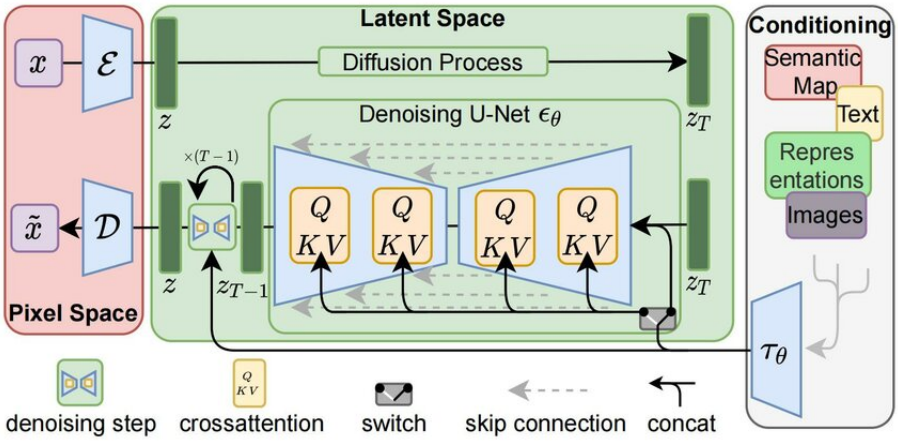


Figura 2.6: Diagrama del modelo de difusión latente propuesto en el artículo [RBL+22]. En él se pueden ver los diferentes elementos que componen el modelo como el mecanismo de atención cruzada, los pasos de reversión de ruido, los procesos de espacio latente y la herramienta de múltiples formatos de datos de entrada.

Hoy en día, los modelos de difusión son el modelo generativo predilecto para la generación de datos sintéticos, desde imágenes hasta audio y vídeo [CTG+24]. Muchas de las herramientas de generación de datos sintéticos más famosas de la actualidad utilizan modelos de difusión, desde Stable Diffusion hasta DALL-E [RDN+22] pasando por Midjourney e Imagen [SCS+22]. Los anteriores mencionados son, a fecha de hoy, herramientas y algoritmos de generación de imágenes sintéticas que implementan la posibilidad de generar imágenes utilizando un texto escrito como entrada ("prompt"), pero hay muchos otros algoritmos, quizás no tan conocidos, para la generación de vídeo como AnimatedDiff [GYR+23], para trabajar con modelos 3D como Shape-E [JN23] y para escribir textos como Diffusion-LM [LTG+22] o GENIE [LGS+23].

2.3.3.1. Stable Diffusion

Stable Diffusion es un modelo generativo de imágenes sintéticas basado en modelos de difusión latentes. Fue desarrollado en 2022 por Stability AI [AI24] a raíz del proyecto de la Universidad de Munich del cual procede el artículo [RBL+22] mencionado previamente. Cuatro de los cinco miembros originales del proyecto se unirían a Stability AI para desarrollar más adelante Stable Diffusion.

Como se ha comentado en las secciones anteriores, Stable Diffusion está compuesto principalmente por un Auto-Encoder Variacional encargado de la transformación del

espacio real al espacio latente, una Red U con una estructura de red neuronal residual (Residual Neural Network (ResNet)) [HZRS16] encargada de la eliminación del ruido y una herramienta para introducir textos como datos de entrada.

Stable Diffusion ha sido entrenado con distintos subconjuntos de imágenes de entrenamiento de LAION [LAI24], utilizando el poder computacional de muchas unidades de procesamiento gráfico dedicado a través de los servicios de Amazon Web Services (AWS) [Ama24].

Stable Diffusion es uno de los modelos de generación de imágenes más famoso y remarcable, y tanto su código como sus pesos son de uso libre y se pueden encontrar en internet. En este proyecto se ha utilizado la versión 2.1 de Stable Diffusion a través de la API de HuggingFace [Sta24].

Capítulo 3

Estado del Arte

El concepto de "deepfake" atañe a una forma avanzada de generar o manipular contenido audiovisual, normalmente mediante el uso de técnicas de inteligencia artificial, con el fin de que este contenido se perciba como contenido real y no sintético [Wes19, RNMS22]. Por ello, la capacidad de discernir entre contenido original y genuino del contenido artificial es de especial relevancia en aspectos como el mantenimiento de la integridad de la información o la protección frente a nuevas amenazas virtuales.

En este capítulo se realizará un repaso de los métodos existentes para la detección de contenido sintético según el origen de este. También se hablará de los conjuntos de datos de entrenamiento que utilizan estos modelos generativos para poder lograr resultados tan prometedores. La Sección 3.1 profundiza en las técnicas de detección específicas para diferentes tipos de modelos generativos, incluyendo GANs y CNNs, mientras que la Sección 3.1.2 se centra en abordar los retos únicos que presentan los modelos de difusión. La Sección 3.2 detalla los conjuntos de datos existentes usados para entrenar estos modelos de detección, subrayando la falta de datos adecuados para los modelos de difusión y cómo esto afecta el desarrollo de tecnologías de detección eficaces. Finalmente, se proporciona un resumen de los artículos más interesantes utilizando una tabla compartiva en la Sección 3.3.

3.1. Detección de Imágenes Sintéticas

A medida que las técnicas de generación de contenido audiovisual sintético avanzan a pasos agigantados, también lo debe hacer aquella tecnología creada para identificar y clasificar estos contenidos, ya que es de gran interés para el ámbito legal, la protección de marcas y contenido propio, la seguridad en las plataformas y redes sociales, la verificación de contenido polémico y mediático, y otros muchos casos donde la generación descontrolada de contenido sintético puede generar consecuencias muy negativas e incluso daños irreparables.

Es necesario que se desarrollen técnicas de detección y clasificación de contenido sintético para poder asegurar la veracidad y transparencia en esta era de la información en la que nos encontramos, particularmente con el aumento exponencial de los datos, información y contenido audiovisual público que existe actualmente y que se encuentra al alcance de todos.

Existen distintas metodologías de detección de información sintética y múltiples

maneras de categorizar estas metodologías. Por ello, se van a diferenciar las distintas herramientas y enfoques de detección según el modelo generativo que haya generado el contenido sintético.

3.1.1. Detección en Redes Generativas Antagónicas y Redes Neuronales Convolucionales

Como se ha visto anteriormente en la Sección 2.1, las Redes Generativas Antagónicas (GANs) han representado hasta hace poco el método por excelencia para la generación y manipulación de imágenes y vídeos sintéticos. Las GANs destacan por su capacidad para producir resultados realistas y creíbles, indistinguibles de contenido real para el ojo humano, generando desafíos significativos en términos de seguridad y verificación de la información. Desde que se popularizó el uso de las GANs, se han estudiado e implementado varias técnicas para detectar contenido generado o manipulado por estas redes neuronales [GCM+21].

Existen distintos enfoques que se pueden aplicar a la hora de querer analizar imágenes para poder clasificarlas como reales o sintéticas. Por ejemplo, en el artículo [YLL19] los autores implementan una herramienta de detección de imágenes sintéticas (siendo algunas de estas generadas por GANs) basada en un clasificador SVM (máquina de soporte vectorial, Support Vector Machine (SVM)) que, entrenando con vectores de características de las poses de las cabezas en las imágenes, es capaz de distinguir si las poses son artificiales o no.

Otros enfoques se centran en que las imágenes producidas por las GANs dejan tras de sí una serie de trazas y artefactos que, aun siendo imperceptibles para los seres humanos, sí que son identificables mediante técnicas y transformaciones.

El artículo [ZKC19] explora varios métodos avanzados para la detección de de imágenes sintéticas generadas por GANs, entrenando clasificadores de imágenes reales frente a imágenes sintéticas, y proporcionan una solución mediante la creación de un simulador de redes GAN al que llamaron "AutoGAN" que simula los artefactos producidos por distintos modelos populares de redes GAN como CycleGAN [ZPIE17] o StarGAN [CCK+18]. AutoGAN realiza un análisis del espectro de frecuencia para detectar artefactos introducidos en el proceso de generación de las GANs (ver Figura 3.1). Entre sus limitaciones destaca que encuentra dificultades a la hora de clasificar imágenes que son producidas por redes GANs cuya estructura sea considerablemente distinta, ya que estas parecen producir otro tipo de artefactos distintos a los producidos por las redes GANs tratadas en el estudio.

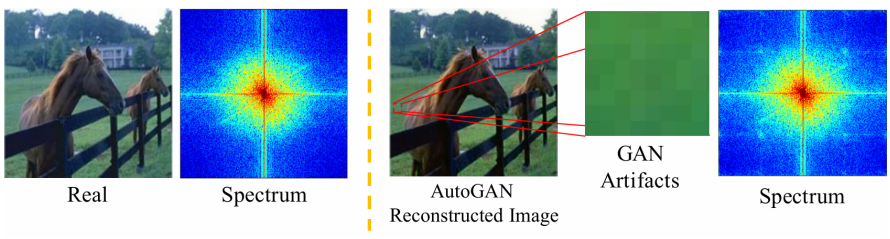


Figura 3.1: Comparativa del espectro de frecuencia entre una imagen real y una generada por el AutoGAN propuesto por el artículo [ZKC19].

Los autores de [WWZ+20b] proponen mejorar el enfoque del anterior artículo mediante el desarrollo de un detector universal capaz de detectar imágenes reales de imágenes sintéticas generadas por cualquier tipo de red neuronal convolucional, independientemente de su arquitectura o conjunto de datos. Para ello, los autores recopilan conjuntos de datos de imágenes generadas por modelos generadores de imágenes como ProGAN [KALL17] y StyleGAN [KLA19] entre otros, siendo entrenados en conjuntos de datos de imágenes como LSUN [YSZ+15] o ImageNet [RDS+15, DDS+09]. El detector de imágenes está basado en un clasificador que utiliza una arquitectura ResNet [HZRS16] y en el análisis de frecuencia de las imágenes mediante la aplicación de distintas transformaciones, parecido al mencionado en [ZKC19]. Las conclusiones del artículo sugieren que las imágenes generadas por redes convolucionales, y en particular las redes GANs, comparten defectos sistemáticos que dejan tras de sí artefactos en las imágenes generadas. Se pueden observar los resultados que los autores obtuvieron resumidos en la Figura 3.2.

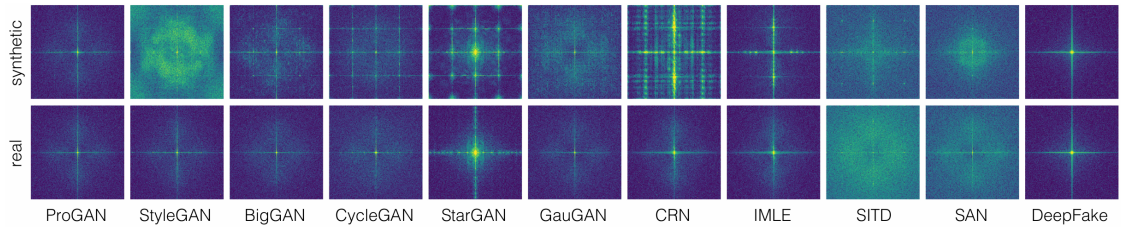


Figura 3.2: Trazas de los artefactos generados por distintas arquitecturas GAN y CNN tras aplicarles transformaciones en el análisis de frecuencia de [WWZ+20b].

Se puede encontrar otro enfoque similar en [GGB20] donde los autores proponen un método basado en el análisis de huellas convolucionales generadas por las GANs. Los autores proponen un algoritmo de maximización de la esperanza al que denominan "Expectation Maximization" para extraer estas características, logrando resultados prometedores en la detección de imágenes sintéticas generadas por redes GANs.

Entre finales del 2019 y mediados del 2020 se realizó una competición diseñada para fomentar el desarrollo de nuevas tecnologías para detectar deepfakes y contenido audiovisual que haya podido ser manipulado o alterado. A esta competición se le dió el nombre de Deepfake Detection Challenge (DFDC) [Ben19], donde se evaluaron propuestas a través de una serie de pruebas y donde se incentivaba a participar a través de una remuneración, siendo la competición apoyada por empresas como Amazon, Facebook y Microsoft.

Entre muchos propósitos, la competición DFDC ha servido para establecer múltiples conjuntos de datos de uso libre para el entrenamiento y desarrollo de herramientas de detección de deepfakes, siendo a fecha de hoy el conjunto de datos de contenido audiovisual sintético más grande disponible públicamente. El artículo [DBP+] proporciona una descripción detallada del conjunto de datos de la DFDC y discute las contribuciones, soluciones y resultados obtenidos a raíz de la competencia generada en la detección de contenido sintético.

3.1.2. Detección de Modelos de Difusión

En la sección anterior se han comentado varias técnicas y metodologías establecidas para identificar imágenes generadas por modelos generativos tradicionales, como las Redes

Generativas Antagónicas (GANs) y otras arquitecturas de redes neuronales. Muchas de estas técnicas han demostrado ser eficientes para la detección y clasificación de manipulaciones y generación de contenido sintético, distinguiendo las imágenes reales de aquellas que han sido generadas de manera artificial o que hayan sido modificadas mediante alguna técnica de deepfake.

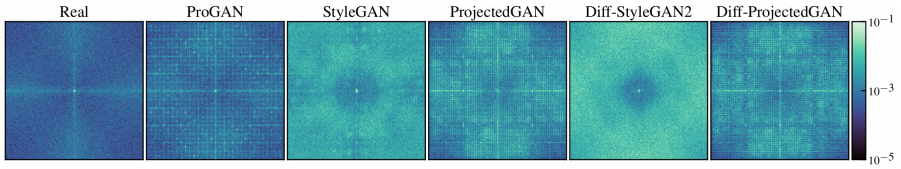
Sin embargo, los modelos de difusión siguen siendo a día de hoy una clase emergente y reciente de modelos generativos. Por ello, las imágenes obtenidas mediante este tipo de modelos presentan dificultades a la hora de ser detectadas y clasificadas. Aunque estos modelos destacan por su capacidad para generar imágenes de alta calidad, el estudio y análisis del contenido generado por estos modelos se encuentra todavía en una etapa temprana e incipiente.

A continuación, se realizará un repaso sobre distintos artículos y estudios que proponen enfoques preliminares para abordar la detección y clasificación del contenido generado por estos modelos de difusión, estando muchos de estos artículos basados en estrategias aplicadas a modelos generativos previos que ya se han revisado en la sección anterior. Cabe destacar que estos enfoques forman parte de un desarrollo emergente y temprano y todavía se requieren investigaciones y aportes en el campo de la detección de contenido sintético generado por los modelos de difusión.

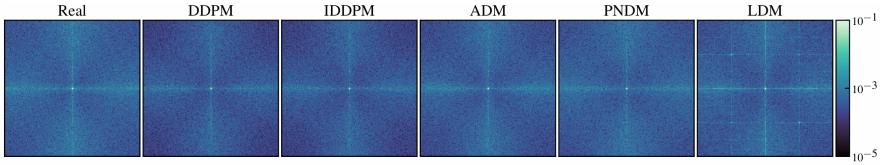
En octubre de 2022 (y revisado por última vez en enero de 2024) se publica uno de los primeros artículos abordando la detección de deepfakes generados por modelos de difusión, que supone uno de los hitos iniciales en este campo. El artículo [RDHF22] analiza la eficacia de los métodos existentes en aquellas fechas para la detección de imágenes sintéticas, diseñados originalmente para modelos generativos anteriores como las redes GANs, y exploran la aplicabilidad de estos métodos a las imágenes producidas por los modelos de difusión. Los resultados del estudio muestran que los detectores basados en redes GANs no son confiables para distinguir entre imágenes reales e imágenes generadas por modelos de difusión pero que, sin embargo, se pueden lograr resultados potenciales al entrenar estos clasificadores con imágenes generadas por modelos de difusión. Esto sugiere que estas últimas imágenes también contienen artefactos producidos en su generación aunque en menor medida que los artefactos producidos por las redes GANs. En el artículo, los autores destacan que los modelos de difusión tienden a subestimar las frecuencias más altas. Se pueden observar las trazas de artefactos producidas por las redes GANs (ver Figura 3.3(a)) en comparación con aquellos artefactos producidos por los DMs (ver Figura 3.3(b)), siendo estos resultados publicados por los autores en el artículo.

En marzo de 2023 se presenta el artículo [GGB23] donde se aborda la evolución de la detección de deepfakes y se presenta un enfoque jerárquico multinivel para mejorar la discriminización entre imágenes reales e imágenes creadas por modelos generativos, utilizando nueves arquitecturas distintas de redes GANs y cuatro arquitecturas de modelos de difusión. La metodología que se propone está basada en el uso de ResNets [HZRS16] concatenadas para realizar una clasificación en distintos niveles, logrando resultados prometedores en la distinción entre redes GANs y DMs.

También en marzo de 2023 se presenta un artículo [WBZ+23] donde se propone un método para detectar imágenes generadas por modelos de difusión basado en un concepto denominado por los autores como “error de reconstrucción por difusión” (Diffusion Reconstruction Error (DIRE)). Este método se basa en calcular el error de reconstrucción entre una imagen como dato de entrada y su contraparte reconstruida por un modelo de difusión pre-entrenado, ya que los autores descubrieron que las imágenes generadas por



(a) Redes GANs.



(b) Modelos de Difusión.

Figura 3.3: Comparativa de las trazas tras aplicar la transformada de Fourier a resultados producidos por distintas arquitecturas de redes GANs y modelos de difusión en [RDHF22].

modelos de difusión pueden ser aproximadamente reconstruidas por los mismos modelos que las crearon.

En abril de 2023 un equipo de investigadores de la Universidad Federico II de Nápoles liderados por Ricardo Corvi presentan el artículo [CCP+23], donde los autores exploran las características distintivas y huellas que dejan las imágenes sintéticas creadas por diferentes modelos generativos como las GANs o los DMs, destacando entre estas huellas lo que ellos denominan "artefactos" en el dominio de Fourier, que hallan aplicando transformaciones a las imágenes y realizando un análisis de frecuencia de éstas.

En mayo 2023 el equipo de Corvi discute en su artículo [CCZ+23] sobre la capacidad de los modelos de difusión para generar medios sintéticos con un impresionante nivel de detalle y por consiguiente el artículo aborda los desafíos asociados con la detección de estos medios. El estudio se centra en el posible impacto de un potencial mal uso sobre la creación de medios falsos con fines maliciosos. Los autores del artículo destacan que los modelos de difusión, aun capaces de generar imágenes de alta calidad, siguen dejando rastros forenses detectables que pueden ser explotados para identificar imágenes sintéticas, incluyendo entre estos posibles rastros inconsistencias en modelación 3D y en las sombras e iluminación (ver Figura 3.4). Por último, comentan que los detectores actuales tienen gran potencial para detectar imágenes generadas por anteriores modelos generativos como las redes GANs, pero que este potencial no se puede aplicar a la detección de contenido sintético generado por los modelos de difusión, incidiendo en la necesidad de desarrollar nuevos métodos y técnicas de detección.

En julio de 2023 los autores del artículo [LDK23] proponen una técnica de detección de imágenes sintéticas creadas por modelos de difusión basada en el análisis de dimensionalidad intrínseca local múltiple, o como ellos la denominan "multiLID" (multi-Logical Intrinsic Dimensionality). Esta técnica fue originalmente desarrollada para discriminar medios sintéticos generados por redes GANs, pero que ha sido capaz de extenderse a los modelos de difusión de manera efectiva y robusta en comparación con los análisis de frecuencia y de los artefactos en el dominio de Fourier, donde los autores comentan las dificultades para diferenciar el contenido sintético generado por los modelos de difusión.

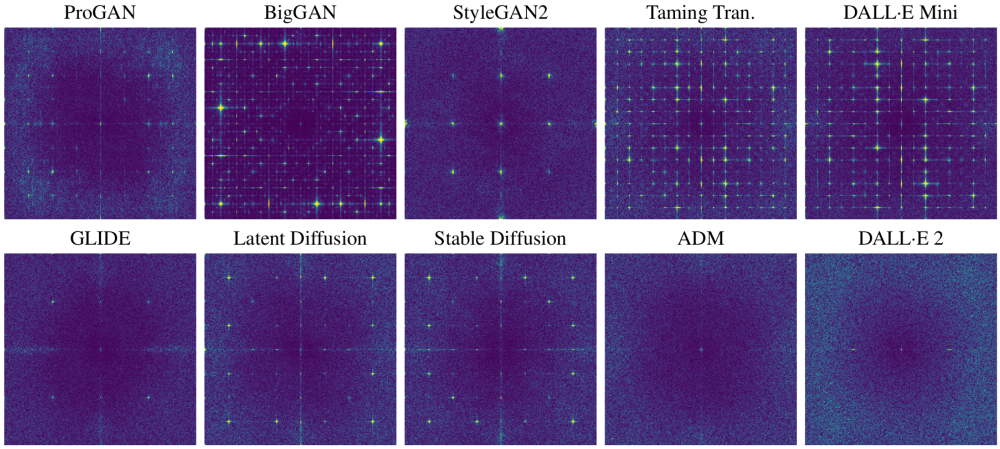


Figura 3.4: Comparativa de las distintas trazas creadas por los artefactos emergentes de la aplicación de la transformada de Fourier para cada tipo de arquitectura mencionada de red GAN y modelo de difusión en [CCZ+23].

En septiembre de 2023 se publica [SHDT23b] donde se analiza la capacidad de generalización de los métodos de detección de deepfakes con un mayor enfoque en los modelos de difusión. El estudio aborda la necesidad de mejores conjuntos de datos compuestos por medios generados por modelos de difusión para el entrenamiento de mejores técnicas de detección de contenido sintético, y propone un conjunto de datos propio que denominaron DeepFakeFace (DFF), compartido de manera libre para apoyar a la comunidad científica en el entrenamiento y evaluación de algoritmos para detectar deepfakes. Este conjunto de datos ha sido creado utilizando un conjuntos de datos de uso libre llamado IMDB-WIKI [RTVG15], y para generar las imágenes han utilizado modelos de difusión latentes como Stable Diffusion v1.5 [RBL+22].

En la Conferencia Internacional sobre Visión por Computadora (International Conference on Computer Vision (ICCV)) [The23] de 2023 se presentó el artículo [AKA23] donde se desarrolla una técnica para la detección de deepfakes generados por modelos de difusión basada en el uso de transformadores de visión (ViTs) combinados con máquinas de soporte vectorial (SVM). El artículo también aborda la escasez de conjuntos de datos abiertos para la detección de deepfakes, por los que los investigadores crearon varios conjuntos de datos personalizados utilizando modelos de difusión como Stable Diffusion [RBL+22] en conjuntos de contenido mediático real como ImageNet [RDS+15, DDS+09], FFHQ [KLA19] y Oxford-IIIT [PVZJ12b].

En la ICCV también se presentó el artículo [FCJ+23] donde se propone otro método para la incorporación de marcas de agua en imágenes creadas por modelos de difusión latentes. Este método, con un enfoque innovador en el campo de detección de contenido sintético para modelos de difusión, consiste en la incrustación de una firma binaria invisible en todas las imágenes generadas por este tipo de modelos utilizando un ajuste particular en el decodificador de los propios modelos, y en el desarrollo de un extractor de marcas de agua que permita recuperar la firma oculta de cualquier imagen. Los autores destacan la utilidad de este enfoque en especial para casos donde se conserva hasta un 10 % de la imagen, mostrando unos resultados con una precisión efectiva.

3.2. Conjuntos de Datos

En la Sección 3.1.1 se ha comentado la disponibilidad de múltiples conjuntos de datos desarrollados para el entrenamiento de modelos generativos convencionales, tales como las Redes Generativas Antagónicas (GANs), las Redes Neuronales Convolucionales (CNNs) y otros modelos, siendo estos recursos fundamentales para el desarrollo de modelos generativos.

Como se ha comentado anteriormente, existe una preocupación y un interés creciente por el desarrollo de técnicas de clasificación, análisis y detección de contenido sintético. Gracias a los propios modelos generativos ya entrenados y a los incentivos y competiciones como la previamente mencionada DFDC [Ben19] también se han desarrollado múltiples conjuntos de datos sintéticos, que son utilizados para entrenar nuevos modelos de detección de contenido sintético.

Sin embargo, se menciona de nuevo que en el ámbito de los modelos de difusión (siendo éstos tecnología de vanguardia en el contexto de la generación de contenido sintético) nos enfrentamos a una marcada escasez de conjuntos de datos públicos. A diferencia de sus predecesores, los modelos de difusión son relativamente nuevos y la comunidad académica y científica no disponen de conjuntos datos generados por modelos de difusión lo suficientemente amplia y robusta. Hasta la fecha, los conjuntos de datos que incluyen imágenes o videos generados por modelos de difusión suelen estar limitados a esfuerzos aislados de grupos de investigación académicos, siendo esto uno de los elementos más limitantes en el desarrollo de herramientas de detección de contenido generado por modelos de difusión.

La falta de conjuntos de datos estandarizados y accesibles implica que los investigadores deben generar sus propios datos para entrenar y probar sus algoritmos de detección, lo cual no solo consume tiempo y recursos, sino que también puede llevar a inconsistencias en la evaluación y comparación de técnicas de detección entre diferentes estudios. En este estudio, se han desarrollado un conjunto de datos específico para el entrenamiento de nuestros modelos de detección, que se comentará con mayor detalle más adelante en esta sección.

3.2.1. Conjuntos de Datos para el Entrenamiento de Modelos Generativos

Los conjuntos de datos que utilizan los modelos generativos para entrenar consisten en miles de archivos audiovisuales, desde imágenes a vídeos, que proporcionan la base necesaria para el aprendizaje y evaluación de éstos modelos. Estos conjuntos de datos contienen una diversa cantidad de características explotables en los modelos generativos a la hora de entrenarlos, desde la variedad y representación equitativa hasta el abordaje de los problemas de sesgo. En este apartado se realizará un repaso de los conjuntos de datos más reconocidos y utilizados en el entrenamiento de los modelos generativos:

- ImageNet [RDS+15, DDS+09]: Definido como un "repositorio de imágenes organizado como una base de datos léxica en el idioma inglés (usando la jerarquía WordNet)" según su página web [FFDR+21], ImageNet es uno de los conjuntos de datos más célebres en la comunidad científica y en el entrenamiento y validación de modelos de aprendizaje profundo. Consta con más de 14 millones de imágenes

anotadas a mano que han sido clasificadas en más de 20.000 categorías.

- LSUN [YSZ+15]: (Large-scale Scene UNderstanding (LSUN)) es un conjunto de datos altamente utilizado en el campo de la visión computacional, especialmente en el entrenamiento y clasificación de escenas. Dos de sus subconjuntos más importantes son LSUN Churches [G.21] y LSUN Bedrooms [Cue23], utilizados en la gran mayoría de los entrenamientos de modelos generativos por su gran amplitud de datos y la calidad de las imágenes.
- MNIST [LCB10]: (Modified National Institute of Standards and Technology database (MNIST)) según su página web [Ult23] es uno de los conjuntos de datos más utilizados en el campo del aprendizaje automático, especialmente en el campo del reconocimiento de imágenes, ya que consta con más de 70.000 imágenes de dígitos escritos a mano (ver Figura 3.5). Ha logrado colocarse como uno de los estándares para el benchmark (o test comparativo) en algoritmos de reconocimiento de imágenes.

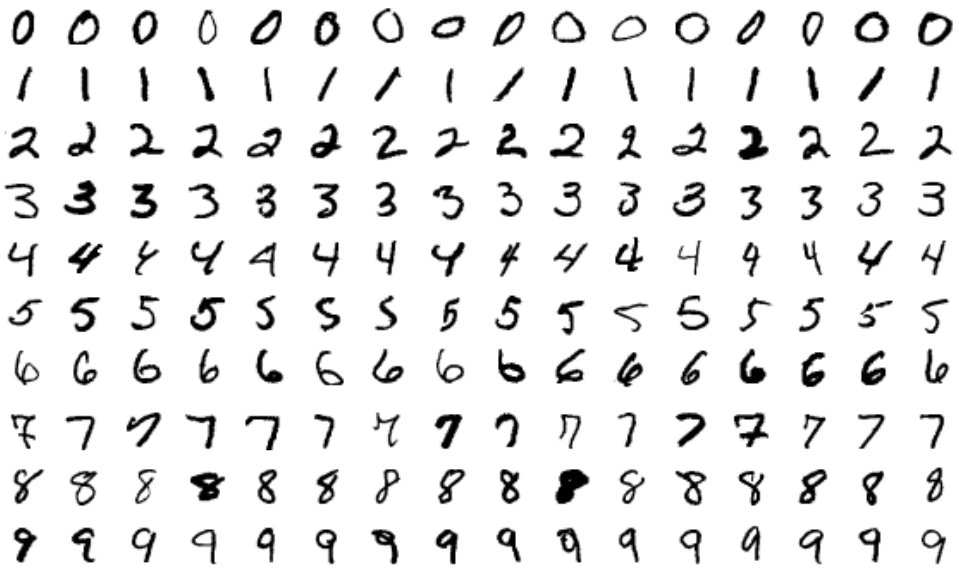


Figura 3.5: Un subconjunto de imágenes de dígitos escritos a mano, procedentes del conjunto de datos de MNIST. Imagen obtenida de la página de [Ult23].

- FFHQ [KLA19]: (Flickr Faces HQ (FFHQ)) es un conjunto de datos compuesto aproximadamente por 70.000 imágenes de rostros humanos en alta resolución (1024x1024), creado específicamente para el entrenamiento y validación de modelos generativos (en particular para redes GANs). Este es uno de los conjuntos de datos de imágenes más utilizados para el entrenamiento de herramientas de deepfake y, por consiguiente, para la detección de este tipo de imágenes falsas.
- CelebA [LLWT15, LLWT21]: (Celebrity Faces Attributes Dataset (CelebA)) es un conjunto de datos que consta de más de 200.000 imágenes de rostros de celebridades y ha sido enormemente utilizado en la comunidad del aprendizaje automático centrado en las tareas relacionadas con el procesamiento de imágenes de rostros faciales.
- MSCOCO [LMB+14, LPR+21]: (Microsoft Common Objects in Context (MSCOCO)) es un conjunto de datos ampliamente reconocido y utilizado en el

campo del aprendizaje automático y la visión computacional. Diseñado con el objetivo de proporcionar recursos de entrenamiento y evaluación de algoritmos para la detección de objetos e instancias, MSCOCO cuenta con más de 330.000 imágenes pertenecientes a 80 categorías distintas, y cada una de ellas cuenta con un etiquetado y anotaciones detalladas.

- LAION [LAI24]: (Large-scale Artificial Intelligence Open Network (LAION)) es una organización centrada en el desarrollo tecnológico dentro del campo del aprendizaje automático y la inteligencia artificial. LAION provee de manera gratuita dos de los conjuntos de datos más importantes y famosos dentro del contexto de la visión computacional, conocidos como LAION-400M [SVB+21] y LAION-5B [SBV+22] por contener aproximadamente 400 millones y 5.000 millones de imágenes respectivamente y que han sido filtradas utilizando CLIP [RKH+21] de OpenAI para asegurar que tanto las imágenes como los textos descriptivos estén alineados.

- Oxford-IIIT Pet Dataset [PVZJ12b]: Según su página web [PVZJ12a], Oxford-IIIT Pet Dataset es un conjunto de datos de mascotas que contiene 7.349 imágenes de 37 razas diferentes de perros y gatos con aproximadamente 200 imágenes por raza, estando cada imagen etiquetada y con anotaciones descriptivas. Fue desarrollado de manera colaborativa entre la Universidad de Oxford y el Instituto Indio de Tecnología de Dehli y es uno de los conjuntos de datos más utilizado para el entrenamiento de imágenes con animales, como se ejemplifica en la Figura 3.6.

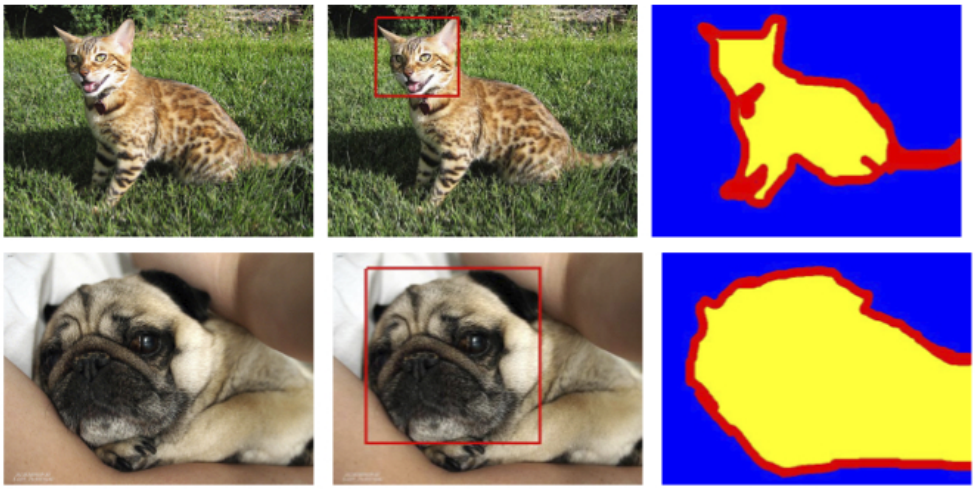


Figura 3.6: Imágenes de perros y gatos obtenidas de [PVZJ12a] donde se ejemplifican las máscaras aplicadas a imágenes del conjunto de datos de Oxford-IIIT Pet Dataset.

- IMDB-WIKI [RTVG15]: El conjunto de datos IMDB-WIKI es, junto con CelebA y FFHQ, uno de los conjuntos de datos más grandes disponibles para el reconocimiento facial y la estimación de edad en imágenes de personas. Provenientes de IMDb y Wikipedia, este conjunto de datos recolecta más de 500.000 imágenes de celebridades de todo el mundo, estando etiquetadas por edad y género además de contener descripciones y datos de las celebridades y anotaciones de metadatos, lo cual lo convierte en uno de los conjuntos de datos por excelencia en el campo del aprendizaje automático y la visión computacional.

3.2.2. Conjuntos de Datos para el Entrenamiento de Modelos de Detección de Imágenes Sintéticas

En la sección anterior se han explorado algunos de los conjuntos de datos más conocidos y utilizados por la comunidad científica para el entrenamiento y validación de los modelos generativos. A continuación, se verán algunos de los conjuntos de datos existentes y de libre uso para el entrenamiento de modelos de detección de imágenes sintéticas, que tienen como objetivo desarrollar y afinar la detección y clasificación de imágenes reales e imágenes generadas de manera artificial. Estos conjuntos de datos han sido producidos por modelos generativos anteriores como redes GANs, CNNs y Auto-Encoders Variacionales entre otros.

- FaceForensics [RCV+18]: es uno de los conjuntos de datos más utilizados en la investigación de la manipulación de imágenes y vídeo, sobre todo en el ámbito de los deepfakes y edición facial. Fue producido en 2018 con el objetivo de desarrollar herramientas que fuesen capaces de combatir contra la desinformación y ser capaz de identificar deepfakes, contando con más de 500.000 fotogramas obtenidos de más de 1.000 vídeos de Youtube que fueron posteriormente manipulados utilizando Face2Face [TZS+16]. En 2019 publicaron una versión mejorada, más robusta y con mayor número de datos llamada FaceForensics++ [RCV+19] (ver Figura 3.7) que es, a fecha de hoy, el conjunto de datos más utilizado para el entrenamiento de herramientas de detección de imágenes sintéticas y deepfakes.

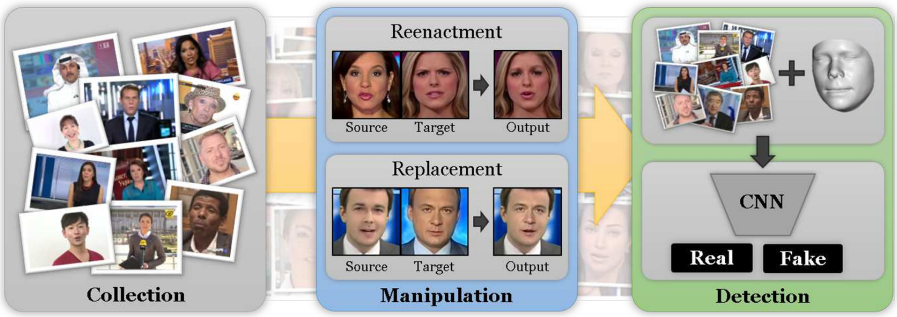


Figura 3.7: Diagrama inicial propuesto en el artículo [RCV+19] sobre el propósito del conjunto de datos FaceForensics++.

- DeeperForensics [JLW+20]: DeeperForensics es el mayor conjunto de datos de detección de falsificación facial más grande hasta la fecha con aproximadamente 60.000 vídeos resultando en un total de 17,6 millones de fotogramas. Los vídeos fueron producidos utilizando un Auto-Encoder Variacional denominado DF-VAE y para ello utilizaron a más de 100 actores de múltiples nacionalidades.

- Celeb-DF [LYS+20]: (Celebrity DeepFake Dataset) es un conjunto de datos compuesto de 590 vídeos reales y 5.639 vídeos manipulados. El objetivo de los autores era proporcionar vídeos manipulados de mayor calidad que el resto de conjuntos de datos y para ello aplicaron un algoritmo de deepfake mejorado que producía menos artefactos que el resto de modelos, y utilizaron vídeos públicos de Youtube, centrándose en particular en entrevistas a famosos.

- DFDC Dataset [DBP+]: Como se ha comentado en la Sección 3.1.1, el DeepFake Detection Challenge fue un concurso patrocinado por Amazon, Facebook y Microsoft

para incentivar la investigación contra el mal uso de los modelos generativos y los deepfake [Ben19]. Uno de los resultados de esta competición fue el conjunto de datos DFDC Dataset, que contiene más de 100.000 vídeos reales y manipulados que han sido etiquetados y clasificados, siendo este conjunto de datos uno de los más usados por la comunidad científica en el campo de la visión computacional.

- WildDeepfake [ZCC+20]: es un conjunto de datos compuesto principalmente por imágenes de caras que han sido manipuladas por técnicas de intercambio de rostros y otro tipo de deepfake. El conjunto está formado por un total de 7.314 secuencias faciales extraidas de 707 vídeos deepfake recolectados de internet.

3.2.2.1. Conjuntos de Datos para el Entrenamiento de Modelos de Detección de Imágenes Sintéticas generadas por Modelos de Difusión

En la sección anterior se han visto los conjuntos de datos más célebres y usados por la comunidad científica para el entrenamiento de modelos de detección de imágenes sintéticas. Pero, como se ha comentado anteriormente en el inicio de la Sección 3.2 , estos conjuntos de datos no contienen imágenes generadas por los modelos más recientes como son los modelos de difusión. Existe una necesidad por parte de la comunidad científica [WBZ+23, CCZ+23, SHDT23b, AKA23] de generar una serie de conjuntos de datos de imágenes sintéticas generadas por modelos actuales para poder entrenar nuevos detectores que ayuden a combatir el mal uso de estas herramientas. A continuación, se comentan recientes propuestas de conjuntos de datos que sí contemplan a los modelos de difusión como generadores de imágenes sintéticas.

- DiffusionForensics [WBZ+23]: en este artículo se propone el uso del conjunto de datos desarrollado por los autores que denominaron "DiffusionForensics". El conjunto de datos contiene imágenes reales de ImageNet, CelebA y LSUN-Bedrooms y modificaciones sintéticas de distintos modelos de difusión entrenados con estos conjuntos de datos mencionados anteriormente. Entre los modelos de difusión más célebres se pueden encontrar modelos como Midjourney, DALL-E [RDN+22] y Stable Diffusion [RBL+22]. Desafortunadamente, a día de hoy el conjunto de datos propuesto por los autores ya no está disponible en el enlace de su proyecto en Github.

- GenImage [ZCY+24a, ZCY+24b]: GenImagen es un conjunto de datos propuesto en el artículo [ZCY+24a] publicado a finales de 2023. En este artículo, los autores exponen que GenImage está compuesto por un total de más de 2.6 millones de imágenes reales e imágenes sintéticas divididas en partes aproximadamente iguales, habiendo sido generadas las imágenes sintéticas por múltiples modelos generativos incluyendo varios modelos de difusión y redes GANs. Los autores enfocaron el artículo a la evaluación de imágenes para su posible clasificación según el tipo de modelo generativo que las hubiese producido y también se centraron en el análisis de imágenes degradadas para poder clasificar aquellas imágenes que se encontrasen en condiciones que no fuesen las ideales para ser analizadas por detectores.

- DeepFakeFace [SHDT23b, SHDT23a]: Como se ha mencionado en la Sección 3.1.2, los autores del artículo [SHDT23b] proponen un conjunto de datos que ellos denominan DeepFakeFace (DFF) compuesto por 120.000 imágenes siendo 30.000 de estas imágenes reales provenientes de IMDB-WIKI [RTVG15] y 90.000 imágenes falsas producidas con Stable Diffusion v1.5 [RBL+22] e InsightFace [DGXZ19]. De

manera similar al artículo [ZCY+24a], los autores se centraron varios métodos, estando el primero de ellos basado en la posible clasificación de imágenes según los distintos modelos generativos que hubiesen podido producir tales imágenes . El segundo método evaluaba el rendimiento del algoritmo de detección para imágenes imperfectas, borrosas o de baja calidad para poder implementar condiciones similares a la de los deepfakes que se pueden encontrar en la realidad.

3.3. Comparativa de Técnicas de Detección

En la Tabla 3.1, se presenta una comparativa de las técnicas de detección de imágenes sintéticas más relevantes en el estado del arte. En ella se resumen los aspectos más importantes de cada propuesta como son los conjuntos de datos, las métricas, los métodos o los algoritmos. Estos artículos engloban tanto aquellos orientados a la detección de imágenes creadas por modelos generativos anteriores como las GANs y CNNs como a la detección de imágenes creadas por los recientes modelos de difusión.

Tabla 3.1: Comparativa de técnicas de detección presentadas en el en el Capítulo 3.

Artículo	**Contexto**	**Técnicas**	**Conjunto de Datos**	**Métricas**	**Conclusiones**
[ZKC19]	Detección y simulación de artefactos en imágenes falsas de GAN	AutoGAN	CycleGAN, ImageNet, MSCOCO	Accuracy	Entrenamiento sin imágenes falsas reales mejora la detección
[WWZ+20b]	Evaluación de detección de imágenes CNN con un detector universal	ProGAN, StyleGAN, BigGAN, CycleGAN, StarGAN	LSUN, ImageNet, MSCOCO, FaceForensics++	Average Precision	Detectores generalizan bien entre modelos CNN
[GGB20]	Detección de deepfakes analizando huellas convolucionales	Algoritmo de maximización de expectativas	Celeb-A	Accuracy	Técnica efectiva para distinguir arquitecturas de GANs
[RDHF22]	Evaluación de la detección de imágenes generadas por modelos de difusión	Detectores de GAN reentrenados	LSUN Bedroom + ADM, LSUN Churches	AUROC, Pd@FAR	Los modelos de difusión son difíciles de detectar, reentrenar mejora
[GGB23]	Detección jerárquica de imágenes generadas por GAN y modelos de difusión	ResNet-34	CelebA, FFHQ, ImageNet	Accuracy	Eficacia superior al 97 % en clasificación por arquitectura
[CCP+23]	Diferencias espectrales entre imágenes reales y sintéticas	GANs, VQ-GANs, DMs, StyleGAN-T	ImageNet, FFHQ, LAION	Espectro de frecuencia radial y angular	Artefactos identificables diferencian imágenes reales de sintéticas
[SHDT23b]	Evaluación de la detección de deepfakes con modelos de difusión	Stable Diffusion v1.5, Stable Diffusion Inpainting, InsightFace	IMDB-WIKI, FaceForensics++, DeepFakeFace	Accuracy, AUC, EER	Variedad en la efectividad de detectores según el método de generación de deepfake
[CCZ+23]	Detección de imágenes falsas usando modelos de difusión	DALL-E 2, Stable Diffusion, GLIDE	COCO, ImageNet, UCID	AUC, Accuracy	Evaluaciones múltiples de detectores de imágenes generadas por modelos de difusión

Los artículos recopilados en la Tabla 3.1 han sido elegidos por la relevancia que tienen

en el ámbito de la detección de imágenes generadas por modelos generativos, que abordan las recientes técnicas de detección aplicadas a modelos de difusión. Estos estudios proponen metodologías innovadoras para identificar imágenes sintéticas y han demostrado resultados positivos. Se puede destacar que la mayoría de ellos basan sus metodologías de detección en el entrenamiento de clasificadores basados en CNNs que usan una arquitectura ResNet, y se destaca que la métrica principal es la precisión (accuracy en inglés). También se observa que utilizan ciertos conjuntos de datos en común como son LSUN, FFHQ, ImageNet y MSCOCO entre otros.

Capítulo 4

Metodología

En este capítulo se procede a explicar qué metodologías y pasos se han seguido para desarrollar nuestro detector de imágenes sintéticas y también se explicarán las decisiones que se han tomado y el por qué de éstas. La Sección 4.1 detalla el uso de clasificadores ResNet adaptados para identificar imágenes generadas por modelos de difusión. En la Sección 4.2, se examinan métodos de análisis del espectro de frecuencia para detectar artefactos en estas imágenes. La Sección 4.3 aborda las técnicas de aumento de datos aplicadas a las imágenes a clasificar. Por último, en la Sección 4.4 se comentan los conjuntos de datos que se valoraron y los que se usaron en el proyecto.

4.1. Clasificadores Binarios de Arquitectura ResNet

En la Sección 3.1.2 ya se han repasado distintos artículos que discrepan unos de otros sobre las posibles técnicas para detectar imágenes generadas por modelos de difusión. Los autores del artículo [RDHF22] exponen que, aunque las técnicas de detección de imágenes sintéticas ya existentes para modelos generativos anteriores (como las que se proponen en [WWZ+20b]) no pueden aplicarse directamente para detectar las imágenes generadas por los más recientes modelos de difusión, estos mismos detectores sí que podrían ser entrenados con nuevos conjuntos de datos actualizados con imágenes generadas por modelos de difusión para que pudiesen detectar estas mismas imágenes. Los autores argumentan que los modelos de difusión siguen dejando trazas de artefactos que pueden ser identificadas si se entrenan correctamente los clasificadores de detección.

Esta teoría también ha sido apoyada por otros grupos de investigación como el de Corvi en sus distintos artículos [CCZ+23, CCP+23] incidiendo en el estudio de las trazas generadas en las imágenes sintéticas por los distintos modelos generativos y la aplicación de transformaciones de Fourier para realizar análisis de frecuencia en estas imágenes. También se pueden considerar los artículos como [LDK23] que exponen que las técnicas como los análisis de frecuencia y los artefactos en el dominio de Fourier no son suficientemente eficaces como para producir resultados concluyentes.

Por ello, en este trabajo se ha propuesto implementar las técnicas previamente introducidas en el artículo [WWZ+20b] para el entrenamiento de detectores de imágenes sintéticas generadas por modelos basados en CNNs pero utilizando conjuntos de datos actualizados con imágenes generadas por modelos de difusión latentes.

En el artículo [WWZ+20b] los autores utilizan un clasificador binario con arquitectura

ResNet-50 [HZRS16] que es entrenado con distintos conjuntos de imágenes producidas por diferentes arquitectura de redes GANs. Además, los autores implementan una serie de preprocesamientos basados en desenfoques y distorsiones a las imágenes previo al entrenamiento de los clasificadores con el objetivo de poder generalizar los análisis de los detectores y poder aplicarlos a datos reales en vez de obtener resultados únicamente con imágenes sintéticas ya preparadas. Esto es de especial interés en el concepto de los detectores de imágenes sintéticas.

Los autores han compartido tanto sus modelos ya entrenados como su código para entrenar nuevos modelos y poder evaluarlos en el Github de su proyecto [WWZ+20a], así que, después de probar los modelos de detección ya entrenados por los autores, el primer paso que se ha realizado para implementar nuestro detector de imágenes sintéticas generadas por modelos de difusión latente ha sido adaptar este código.

Para ello se han implementado modificaciones en la arquitectura inicial de los clasificadores debido a que los modelos que utilizan los autores entrenan múltiples categorías a la vez, como pueden ser personas, distintos animales, objetos o localizaciones, mientras que en esta investigación únicamente se tiene interés en poder clasificar rostros humanos. También se modificó el código original para poder utilizar distintas arquitecturas ResNet tanto en los procesos de entrenamiento y validación como en los procesos de prueba.

4.1.1. Arquitecturas ResNet en este proyecto

Se han probado tres de las arquitecturas ResNet [HZRS16] más célebres y utilizadas en el campo de la visión computacional: la ResNet50, la ResNet34 y la ResNet18. Cada una de estas estructuras contiene un número de capas ocultas igual al número que contiene su nombre.

- ResNet50: Esta es la arquitectura más compleja que se ha utilizado, siendo conocida por su uso en tareas costosas o en el trabajo con conjuntos de datos variados y grandes, ya que permite modelar relaciones mucho más complejas que el resto de capas que se han usado. Esta arquitectura utiliza un tipo de bloques de capas convolucionales conocido como Bottleneck, compuesto por una capa 1x1, una 3x3 y otra 1x1 al final, con el objetivo de operar con espacios dimensionales reducidos para aumentar la eficiencia. La arquitectura ResNet50 está dividida en cuatro secciones de procesamiento de características y cada una de estas secciones contiene un número distinto de bloques, siendo la primera sección de tres bloques, la segunda de cuatro bloques, la tercera de seis bloques y la última de tres bloques.
- ResNet34: Esta es la arquitectura intermedia, la cual tiene menos capas ocultas que la ResNet50 pero sigue manteniendo un alto grado de complejidad para poder realizar tareas de visión computacional. Esta arquitectura un tipo de bloques de capas convolucionales conocido com BasicBlock, compuesto por dos capas con kernels de 3x3 cada capa. Es una estructura más simple que la de Bottleneck y por ende utilizada en variantes menos profundas de ResNet. La arquitectura ResNet34 mantiene la misma distribución de secciones y bloques que la ResNet50, aunque las secciones están compuestas por bloques de tipo BasicBlock en vez de bloques de tipo Bottleneck.
- ResNet18: Esta es la arquitectura más simple que se ha decidido usar en el proyecto. Este tipo de arquitecturas destaca por su eficiencia en el uso de memoria y por

su velocidad en comparación con otro tipo de arquitecturas ResNet, aunque es posible que tenga mayores dificultades cuando se trate de trabajar con conjuntos de datos más grandes y complejos. Este tipo de arquitectura también utiliza los mismo bloques BasicBlock que se ha visto en la arquitectura ResNet34, aunque contiene una distribución de secciones distinta, conteniendo también cuatro secciones con dos bloques por cada sección.

4.2. Análisis del Espectro de Frecuencia

En la sección anterior se han comentado algunos de los artículos que exponen que las imágenes generadas por modelos de difusión siguen generando trazas de artefactos, y que estas trazas se pueden detectar e identificar utilizando distintos análisis de frecuencia.

En este trabajo se ha implementado el análisis del espectro de frecuencia que proponen en el artículo [RDHF22] ya que en él los autores realizan varios estudios y comparaciones de imágenes generadas por distintos modelos generativos (incluyendo modelos de difusión) y distintos conjuntos de datos. Estas comparaciones resultan muy útiles para comprobar la existencia de las supuestas trazas que los distintos modelos generativos dejan como rastros inicialmente imperceptibles en las imágenes que generan.

Se ha elegido implementar el análisis de tres transformaciones y el análisis de densidad del artículo [RDHF22] que se puede aplicar a nuestros conjuntos de imágenes para poder realizar el respectivo análisis. Estas transformaciones son la Transformada de Fourier Discreta (Discrete Fourier Transform (DFT)), su versión con Filtro de Paso Alto (High Pass - Discrete Fourier Transform (HP-DFT)), la Transformada de Coseno Discreta (Discrete Cosine Transform (DCT)) y el Cálculo del Espectro de Densidad (Spectrum Density). A continuación, se verá en qué consiste cada una de estas transformaciones.

4.2.1. Transformada de Fourier Discreta

La transformada de Fourier discreta es comúnmente utilizada para el análisis de señales y procesamiento de imágenes porque permite transformar una imagen del dominio espacial al dominio de frecuencia. El objetivo de este tipo de transformaciones es poder identificar las posibles trazas de artefactos que generan los modelos de difusión latentes. La DFT de una imagen en escala de grises se define según la Ecuación (4.1).

$$I_{\mathrm{DFT}}[k,l] = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} I[x,y] \exp^{-2\pi i \frac{xk}{H}} \exp^{-2\pi i \frac{yl}{W}} \tag{4.1}$$

Ecuación 4.1: Transformada de Fourier Discreta para escala de grises, descrita en los anexos de [RDHF22].

donde:

- $I[x,y]$ representa el valor de la imagen I en el píxel ubicado en la posición (x,y).
- H y W denotan la altura y la anchura de la imagen I, respectivamente.

- k y l son los índices de frecuencia correspondientes a las dimensiones vertical y horizontal de la imagen transformada.
- i es la unidad imaginaria, y exp denota la función exponencial.

4.2.2. Filtro de Paso Alto aplicado a la Transformada de Fourier

El filtro de paso alto es utilizado para destacar detalles en la alta frecuencia de una imagen con el objetivo de enfocar características como bordes y texturas finas. La aplicación de un filtro de paso alto tiene lugar antes de realizar la transformada de Fourier. La transformada de Fourier con filtro de paso alto se define según la Ecuación (4.2).

$$I_{\text{HP-DFT}}[k,l] = \sum_{x=0}^{H-1}\sum_{y=0}^{W-1}(I[x,y] - I_{\text{med}}[x,y])\exp^{-2\pi i\frac{xk}{H}}\exp^{-2\pi i\frac{yl}{W}} \tag{4.2}$$

Ecuación 4.2: Transformada de Fourier Discreta con Filtro de Paso Alto (esta ecuación no se encuentra en el artículo [RDHF22] pero se deduce del comportamiento del código compartido por los autores).

donde:

- $I[x,y]$ es el valor original del píxel en la imagen.
- $I_{\text{med}}[x,y]$ es el valor del píxel después de aplicar un filtro mediano, utilizado para suavizar la imagen y extraer la componente de baja frecuencia.
- H y W denotan la altura y la anchura de la imagen I, respectivamente.
- k y l son los índices de frecuencia correspondientes a las dimensiones vertical y horizontal de la imagen transformada.
- i es la unidad imaginaria, y exp denota la función exponencial.

4.2.3. Transformada de Coseno Discreta

La transformada de coseno discreta se utiliza en el procesamiento de señales y compresión de imágenes, ya que funciona muy bien para concentrar la energía de la señal en unos pocos coeficientes. La DCT de una imagen en escala de grises se define según la Ecuación (4.3).

$$I_{\text{DCT}}[k,l] = \sum_{x=0}^{H-1}\sum_{y=0}^{W-1} I[x,y]\cos\left[\frac{\pi}{H}(x+\frac{1}{2})k_x\right]\cos\left[\frac{\pi}{W}(y+\frac{1}{2})k_y\right] \tag{4.3}$$

Ecuación 4.3: Transformada de Coseno Discreta para imágenes en escala de grises, descrita en los anexos de [RDHF22].

donde:

- $I[x, y]$ representa el valor de la imagen I en el píxel ubicado en la posición (x, y).
- H y W denotan la altura y la anchura de la imagen I, respectivamente.
- k y l son los índices de frecuencia correspondientes a las dimensiones vertical y horizontal de la imagen transformada.

4.2.4. Cálculo del Espectro de Densidad

La densidad espectral de una imagen, calculada a partir de su espectro de Fourier, se define según la Ecuación (4.4).

$$\hat{S}(r) = \frac{1}{2\pi} \int_0^{2\pi} S(r, \theta)\, d\theta; \qquad r = \sqrt{\frac{k^2 + l^2}{\frac{1}{4}\,(H^2 + W^2)}}, \quad \theta = \arctan 2(k, l) \tag{4.4}$$

Ecuación 4.4: Cálculo del espectro reducido, descrito en los anexos de [RDHF22].

donde:

- $S(r, \theta)$ representa el espectro de potencia en coordenadas polares, con r y θ indicando el radio y el ángulo, respectivamente.
- $\hat{S}(r)$ es el espectro de potencia promediado azimutalmente (densidad espectral).

El espectro de potencia $S(r, \theta)$ se obtiene del cuadrado de la magnitud de los coeficientes de la Transformada de Fourier: $S[k, l] = |I_{DFT}[k, l]|^2$.

La frecuencia máxima se obtiene a partir de la frecuencia de Nyquist $f_{nyq} = \sqrt{k^2 + l^2} = H/\sqrt{2}$ para una imagen cuadrada donde $H = W$.

4.3. Técnicas de Aumento de Datos

En el artículo [WWZ+20b], los autores describen las opciones de distorsión de imágenes que se le pueden aplicar al conjunto de datos en su entrenamiento, con el objetivo de lograr un entrenamiento más generalista y que el aprendizaje de los modelos sea capaz de extender su discriminación más allá del conjunto de datos de entrenamiento y validación. Los autores describen distintas técnicas de alteración de imágenes que se aplican al conjunto de imágenes antes de ser recortadas en el proceso, configurables mediante parámetros en el entrenamiento.

- Desenfoque Gaussiano: Consiste en aplicar un desenfoque gaussiano a las imágenes con una intensidad que varía aleatoriamente dentro de un rango definido (un valor uniforme entre 0 y 3). Esta técnica suaviza la imagen al usar una función matemática gaussiana para transformar los píxeles, lo cual puede ayudar a que el modelo sea más robusto ante variaciones menores en las imágenes de entrada y reducir el sobreajuste durante el entrenamiento.

- Compresión JPEG: Esta técnica se aplica con una probabilidad definida por parámetro (en esta investigación se ha establecido al 50 %) y utiliza las bibliotecas OpenCV y Python Imaging Library (PIL) para comprimir y luego descomprimir las imágenes. La calidad de la compresión está establecida en un intervalo aleatorio entre 30 y 100 (donde 100 es la máxima calidad y menor compresión). Al entrenar con imágenes JPEG, el modelo puede aprender a ser menos sensible a estos artefactos, lo cual es útil cuando se trata de imágenes que podrían ser comprimidas en escenarios del mundo real.

4.4. Conjuntos de Datos

En la Sección 4.1 se han repasado varios enfoques encontrados en diversos artículos recientes del campo de investigación de la detección de imágenes sintéticas generadas por modelos de difusión, y se ha expuesto que uno de los objetivos de este proyecto sería el de intentar entrenar clasificadores binarios con arquitectura ResNet [HZRS16] para que pudiesen discriminar imágenes reales de imágenes generadas por modelos de difusión latente. Por ello, es necesario disponer de un conjunto de datos con imágenes reales y otro conjunto de datos con imágenes generadas por estos modelos de difusión, y los conjuntos deben ser lo suficientemente grandes y de gran calidad como para poder realizar el entrenamiento del clasificador.

4.4.1. Conjuntos de Datos Propio

Durante el desarrollo del proyecto, uno de los retos a resolver fue el de encontrar conjuntos de datos que sirviesen para entrenar los clasificadores debido a los problemas comentados en la Sección 3.2.2.1. Tras no encontrar conjuntos de datos válidos para el entrenamiento de clasificadores, se valoraron distintas opciones como plantear otro tipo de estrategias de detección.

Finalmente, se propuso la idea de desarrollar un conjunto de datos propio que fuese válido para el entrenamiento de clasificadores y otras técnicas de detección de imágenes sintéticas generadas por modelos de difusión, como habían hecho otros autores en sus proyectos. Una de las intencionalidades de este conjunto de datos era la de centrarse en la detección de rostros sintéticos específicamente, y también la de generar imágenes con las versiones más recientes de Stable Diffusion [RBL+22]. Así pues, se acabó desarrollando un conjunto de datos con imágenes sintéticas desarrolladas con Stable Diffusion (ver Figura 4.1).

Mientras se realizaban las primeras pruebas prácticas con el conjunto de datos propio, se continuó con la investigación en el campo de la detección de imágenes sintéticas de manera paralela, y así se encontró el artículo [SHDT23b] mencionado previamente en la Sección 3.2.2.1. Los autores del artículo ofrecen el conjunto de datos DFF que también se acaba utilizando en este proyecto (como se comentará en la Sección 4.4.2) junto con el conjunto de datos propio.

Por último, se ha querido proponer otro enfoque a la hora de compartir el conjunto de datos, y es que se ha desarrollado un simple script de Python para generar imágenes con Stable Diffusion v2.1, utilizando la API de HugginFace [Sta24]. Este script permite elegir mediante parámetros varias configuraciones a la hora de generar las imágenes, como el

Figura 4.1: Ejemplos de imágenes generadas sintéticamentes utilizando Stable Diffusion v2.1 pertenecientes al conjunto de datos propio.

número de las imágenes, y además cuenta con un prompt dinámico que tiene como objetivo añadir una gran variedad de posibilidades en la generación de las imágenes del conjunto de datos. Gracias a esto, cualquier persona puede generar conjuntos de datos propios y adaptables a las necesidades que requieran, sin tener que lidiar con los inconvenientes de tener que descargar todas las imágenes de proyectos ajenos.

Se detalla más acerca de en qué consiste y cómo funciona el script `dataset_generator.py` en el Anexo A.

4.4.2. Conjuntos de Datos Externos

En las Secciones 3.1.2 y 3.2.2.1 se han repasado varios artículos donde se comentaba la importancia y la necesidad de la comunidad científica de generar conjuntos de datos que contuviesen imágenes generadas por modelos de difusión para el posible desarrollo de herramientas y técnicas de detección de imágenes sintéticas. Algunos de estos artículos proponen sus propios conjuntos de datos en sus proyectos como aportación a la comunidad científica, normalmente incluyendo enlaces a sus proyectos en los propios artículos.

Para la realización de este proyecto se fueron probando estos posibles conjuntos de datos pero debido a varios motivos, estos fueron descartados. Entre los motivos encontramos la discontinuidad de los conjuntos de datos en la nube, archivos corruptos e imágenes que no consistían en rostros faciales.

Como se ha comentado en la sección anterior, se decidió entonces generar un conjunto de datos propio usando Stable Diffusion v2.1 [RBL+22], y durante el avance del proyecto se dió con el conjunto de datos DFF, propuesto en el artículo [SHDT23b] en septiembre de 2023. Este conjunto de datos está compuesto por los siguientes subconjuntos de imágenes:

- 30.000 imágenes reales provenientes del conjunto de datos IMDB-WIKI [RTVG15]
- 30.000 imágenes sintéticas generadas con la herramienta de deepfakes InsightFace [DGXZ19].
- 30.000 imágenes sintéticas generadas con Stable Diffusion v1.5 [RBL+22].
- 30.000 imágenes sintéticas generadas con Stable Diffusion v1.5 usando Inpainting.

El conjunto de imágenes DFF cumple con las características necesarias para poder entrenar los clasificadores binarios centrados en rostros humanos y por ende se decidió implementar este conjunto de imágenes en el proyecto junto con el conjunto de datos propio comentado en la sección 4.4.1.

Capítulo 5

Experimentos y Resultados

En este capítulo se detallarán las pruebas y resultados obtenidos a lo largo del estudio, donde se examina la eficiencia y desempeño de los distintos clasificadores, entrenados con diferentes parámetros y formatos en sus estructuras neuronales, buscando un enfoque con el que hallar las configuraciones más eficaces y eficientes. Los clasificadores también han sido entrenados con diferentes conjuntos de datos con el objetivo de comparar también la eficiencia de éstos tanto en el entrenamiento como en la evaluación posterior. También se comentará acerca de los distintos análisis de frecuencia que se han realizado para examinar los conjuntos de datos con los que se ha trabajado.

Adjuntamos un enlace al repositorio donde encontrar el código del proyecto en [CGdMG24].

La Sección 5.1 explora cómo diversas configuraciones de ResNet afectan el rendimiento en la clasificación de imágenes sintéticas y el impacto de las técnicas de aumento de datos. La Sección 5.2 muestra las evaluaciones de los modelos con diferentes conjuntos de datos, destacando sus capacidades y limitaciones. La Sección 5.3 analiza la eficacia de los modelos en entornos de prueba, enfocándose en su capacidad para generalizar y detectar imágenes sintéticas. Por último, la Sección 5.4 estudia las características espectrales de los conjuntos de datos mediante técnicas avanzadas de análisis de frecuencia, identificando diferencias clave entre imágenes reales y sintéticas.

5.1. Entrenamiento de Modelos

Uno de los objetivos principales de este proyecto era el de entrenar distintas redes neuronales convolucionales con el objetivo de que aprendiesen a discriminar imágenes reales de imágenes generadas por modelos de difusión latentes. Para ello, se han entrenado distintas redes neuronales de arquitectura ResNet utilizando una versión modificada del código propuesto en el artículo [WWZ+20b], utilizando estructuras distintas con el fin de diferenciar cuáles de ellas es la más adecuada para este tipo de clasificadores.

5.1.1. Conjuntos de Datos y Distribución

Las imágenes que forman los distintos conjuntos de datos utilizados (nuestro propio conjunto y las distintas categorías que formarían el conjunto de imágenes DFF [SHDT23b])

se categorizan como "0_real" para las imágenes reales y "1_fake" para las imágenes generadas de manera sintética.

Nuestro conjunto de imágenes está compuesto por un total de 20.000 imágenes, siendo la primera mitad imágenes reales provenientes del conjunto de datos FFHQ [KLA19] y la segunda mitad imágenes sintéticas utilizando descripciones (promps) para generar imágenes con Stable Diffusion 2.1 [RBL+22].

El conjunto de imágenes DeepFaceFake propuesto en el artículo [SHDT23b] está dividido en 30.000 imágenes reales provenientes del conjunto de datos IMDB-WIKI [RTVG15], 30.000 imágenes sintéticas generadas por la herramienta InsightFace [DGXZ19], 30.000 imágenes sintéticas generadas utilizando la técnica de inpainting comentada en la Sección 5.2.4 usando Stable Diffusion 1.4 [RBL+22] y 30.000 imágenes sintéticas utilizando descripciones (promps) para generar imágenes con Stable Diffusion 1.4 [RBL+22]. Por lo tanto, se ha decidido entrenar distintos modelos utilizando estos conjuntos de imágenes y la combinación de estos para poder calificar la eficacia del entrenamiento.

5.1.1.1. Distribución del Entrenamiento en Redes Neuronales

Respecto a la distribución de cada uno de estos conjuntos de datos, se ha decidido mantener la distribución propuesta por en [WWZ+20b] de imágenes entre las tres colecciones que formarían el proceso completo de entrenamiento de estos clasificadores, siendo estas el conjunto "train" con un total del 70 % de las imágenes del conjunto de datos entero, el conjunto "val" (validation) que comprendería el 15 % y el conjunto "test" que comprendería el 15 % restante.

- Entrenamiento (train): Son los datos sobre los cuales el modelo se entrena activamente, cuyas etiquetas o respuestas correctas están disponibles para el modelo y éste las usa para ajustar sus parámetros y pesos internos con el objetivo de minimizar el error de predicción.
- Validación (val): Estos datos se utilizan para proporcionar una evaluación del ajuste del modelo durante el entrenamiento, afectando a parámetros como la tasa de aprendizaje o para decidir cuándo detener el entrenamiento y evitar el sobreajuste (overfitting).
- Prueba (test): Son datos que el modelo nunca ha visto durante el entrenamiento y que se utilizan después de que el modelo ha sido entrenado y validado para probar su rendimiento. Es la prueba final para evaluar la generalización y la eficacia del modelo en condiciones reales o desconocidas.

5.1.2. Métricas del Entrenamiento

En este trabajo se analizará la evolución del entrenamiento de los modelos a través de métricas comúnmente utilizadas en el entrenamiento de clasificadores como son la precisión, la precisión media y la función de pérdida.

- Precisión (Accuracy): La precisión mide el porcentaje de predicciones correctas realizadas por el modelo respecto al total de predicciones. Se calcula según la Ecuación (5.1).

$$\text{Precisión} = \frac{\text{Número de predicciones correctas}}{\text{Total de predicciones}} \tag{5.1}$$

- Precisión Media (Average Precision, AP): La precisión media es una métrica que evalúa la precisión del modelo en función de la tasa de verdaderos positivos (recall en inglés). La precisión media resume la curva precisión-recall, que es una gráfica que muestra la precisión del modelo frente al recall para diferentes umbrales. Se calcula según la Ecuación (5.2).

$$AP = \int_0^1 p(r)\, dr \tag{5.2}$$

  donde p es la precisión y r es el recall.

- Función de Pérdida (Loss): La función de pérdida cuantifica cuánto se desvían las predicciones del modelo de los valores reales. Al ser una tarea de clasificación se utilizará la función de pérdida de entropía cruzada (Cross-Entropy Loss), que mide la diferencia entre las distribuciones de probabilidad real y la predicha por el modelo. Se calcula según la Ecuación (5.3).

$$\text{Pérdida} = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{5.3}$$

  donde $y_{o,c}$ es un indicador binario que es 1 si la clase c es la correcta para la observación o, y 0 en caso contrario; $p_{o,c}$ es la probabilidad predicha por el modelo de que la observación o pertenece a la clase c; y $\sum_{c=1}^{M}$ es la suma sobre todas las clases M, asegurando que todas las clases sean consideradas.

5.2. Resultados y Comparaciones en los Entrenamientos

Se ha decidido nombrar a los modelos de clasificación en función del conjunto de datos con el que han sido entrenados seguido del número de arquitectura ResNet que utilizan. Así pues, un clasificador que utilice el conjunto de datos que hemos desarrollado nosotros y una arquitectura ResNet18 se llamará 'own18', mientras que un clasificador que utilice el conjunto de datos de InsightFace dentro del conjunto de datos de DFF con una arquitectura ResNet50 se llamará 'insight50'. A continuación, se hará un repaso de las pruebas y comparaciones que se han realizado durante los entrenamientos de los distintos modelos.

Cabe mencionar que para el entrenamiento de los siguientes modelos se ha utilizado un ordenador con 32GB de RAM, un procesador AMD Ryzen 5 5600X y una unidad de procesamiento gráfico de NVIDIA GeForce RTX 3080 Ti. Para el entrenamiento de los modelos se ha implementado CUDA de NVIDIA para habilitar el entrenamiento con la unidad de procesamiento gráfico.

Por cuestiones de eficiencia se comenta que, durante los entrenamientos, los modelos con arquitectura ResNet18 utilizaban entorno a un 40 % de la capacidad máxima de la unidad de procesamiento gráfico, mientras que los modelos de arquitectura ResNet34 utilizaban entorno al 60-70 % y los modelos con arquitectura ResNet50 utilizaban entre el 95 % y el 100 % de la capacidad máxima de la unidad de procesamiento gráfico.

Los tiempos de entrenamiento para cada modelo estaban determinados por el número de pasos en cada entrenamiento, realizando aproximadamente 20.000 pasos por hora de entrenamiento. Así, los modelos que superan los 35.000 pasos en sus entrenamientos solían terminar tras 1 hora y 45 minutos aproximadamente.

5.2.1. Entrenamientos con y sin Parámetros de Aumento

La primera prueba que se ha realizado ha sido comparar los resultados de entrenar los clasificadores con nuestro propio conjunto de datos pero utilizando tanto distintas arquitecturas ResNet como la aplicación de los parámetros de aumento comentados en la Sección 4.3, en la cual se ha comentado que el objetivo de los parámetros es el de aplicar de manera aleatoria entre algunas fotos del conjunto de datos de entrenamiento y validación cierto grado de desenfoque gaussiano y cierto grado de compresión con el objetivo de promover la generalización del clasificador a la hora de discriminar imágenes a coste de perder eficacia y precisión en el proceso de clasificación.

Como se puede ver en los resultados de la Figura 5.1, el hecho de aplicar estos parámetros no supone un sacrificio en la precisión, puesto que los resultados obtenidos por ambos clasificadores son prácticamente idénticos. Ambos obtienen resultados de precisión y precisión media por encima de 0.99, concordando estos resultados con las discusiones presentadas en el artículo [WWZ+20b] donde se expresa que el uso de este tipo de técnicas de aumento es favorable y recomendable en tareas de clasificación.

5.2.2. Modelos Basados en el Conjunto de Datos Propio 'own'

A continuación, se realizarán comparaciones entre los modelos que han sido entrenados utilizando el conjunto de datos que hemos desarrollado, presentado en la Sección 4.4, compuesto por 10.000 imágenes generadas utilizando Stable Diffusion v2.1 [RBL+22] y 10.000 imágenes del conjunto de imágenes reales de FFHQ [KLA19]. Se han entrenado tres modelos diferentes: 'own18', 'own34' y 'own50'.

Se puede observar con los resultados en la Figura 5.2 que, tanto el modelo entrenado con arquitectura ResNet18 como los modelos entrenados con arquitecturas ResNet34 y ResNet50, obtienen resultados muy similares. Alcanzan una puntuación en la precisión de más de 0.99 y de una puntuación de precisión media cercana al 1.

Se destaca que los resultados obtenidos han alcanzado un número de pasos de entrenamiento (steps) determinado por un mecanismo conocido como detención prematura (early-stopping) con el objetivo de evitar el sobreajuste (overfitting), y que el modelo con arquitectura ResNet34 es aquel que menos pasos ha necesitado, seguido del modelo con arquitectura ResNet50 y por último el modelo con arquitectura ResNet18.

5.2.3. Modelos Basado en el Conjunto de Datos 'text2img' de DeepFakeFaces

Como se ha mencionado en la Sección 4.4.2, el conjunto de datos text2img está compuesto por 30.000 imágenes generadas utilizando Stable Diffusion v1.4 [RBL+22], y se ha utilizado el conjunto de imágenes reales propuesto también por DFF que consiste en 30.000 imágenes del conjunto de datos IMDB-WIKI [RTVG15]. Se han entrenado 3 modelos: 'text2img18', 'text2img34' y 'text2img50'.

(a) Clasificadores con arquitectura ResNet18.

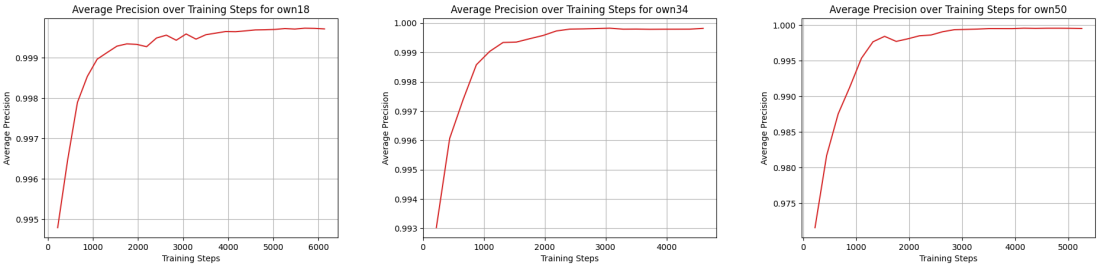
(b) Clasificadores con arquitectura ResNet34.

(c) Clasificadores con arquitectura ResNet50.

Figura 5.1: Comparación entre clasificadores con y sin parámetros de aumento aplicados. En rojo se aprecia el clasificador con los parámetros de aumento aplicados. En azul, el clasificador sin los parámetros de aumento. Para cada figura, las gráficas de la izquierda corresponden con la métrica de precisión, las gráficas centrales corresponden con la precisión media y las gráficas de la derecha corresponden con la función de pérdida.

Como se puede observar en la Figura 5.3, los resultados obtenidos utilizando distintas arquitecturas ResNet para este conjunto de imágenes son muy similares entre ellos, alcanzando resultados muy precisos en aproximadamente el mismo número de pasos de entrenamiento (25.000). Aunque las puntuaciones de precisión logran marcas similares, el número de pasos se acrecienta en comparación con el entrenamiento de los modelos entrenados con nuestro conjunto de datos, y esto se debe principalmente a que el conjunto de datos 'text2img' es tres veces más grande que el nuestro.

(a) Métrica de precisión (accuracy) durante el entrenamiento.



(b) Métrica de precisión media (average precision) durante el entrenamiento.

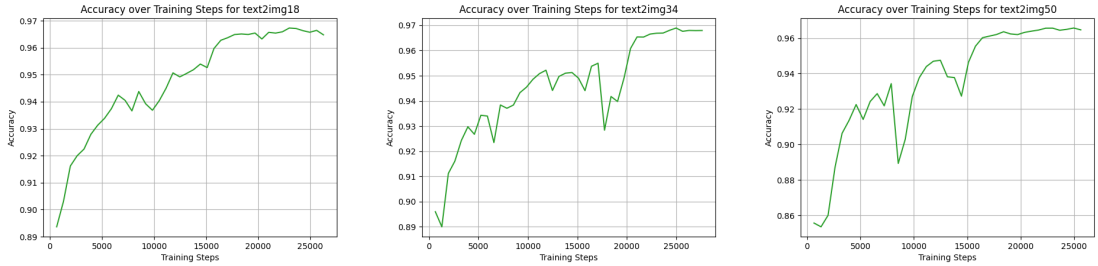
(c) Métrica de función de pérdida (loss) durante el entrenamiento.

Figura 5.2: Comparación entre clasificadores con distinta arquitectura ResNet, entrenados con el conjunto de datos propio 'own'.

5.2.4. Modelos Basado en el Conjunto de Datos 'inpainting' de DeepFakeFaces

El conjunto de datos 'inpainting' está compuesto por 30.000 imágenes donde se ha implementado esta técnica descrita en la Sección , utilizando también Stable Diffusion v1.4 [RBL+22] para ello. Para el conjunto de datos reales se han utilizado las 30.000 imágenes reales de IMDB-WIKI [RTVG15] como los entrenamientos de los modelos anteriores.

Se puede observar, si se comparan las Figuras 5.4(a), 5.4(b) y 5.4(c), que el hecho de utilizar distintas arquitecturas ResNet en estos modelos sí que tiene un impacto tanto en su entrenamiento como en su puntuación de precisión final. El modelo 'inpainting18' logra superar una marca de precisión de 0.9 pero para ello casi alcanza 35.000 pasos de entrenamiento, mientras que el resto de arquitecturas no logra llegar a 0.9 y detiene su entrenamiento antes, entre los 22.000 pasos para 'inpainting34' y 28.000 pasos para 'inpainting50' aproximadamente. Se observa que, aunque logren buenas puntuaciones de precisión, estos modelos tienen mayores dificultades para discriminar imágenes reales de sintéticas que los modelos comentados anteriormente.



(a) Métrica de precisión (accuracy) durante el entrenamiento.

(b) Métrica de precisión media (average precision) durante el entrenamiento.

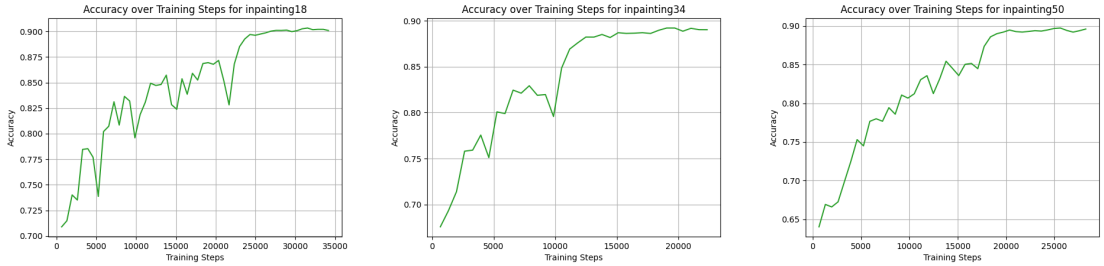
(c) Métrica de función de pérdida (loss) durante el entrenamiento.

Figura 5.3: Comparación entre clasificadores con distinta arquitectura ResNet, entrenados con el conjunto de datos 'tex2img' del DeepFakeFace [SHDT23b].

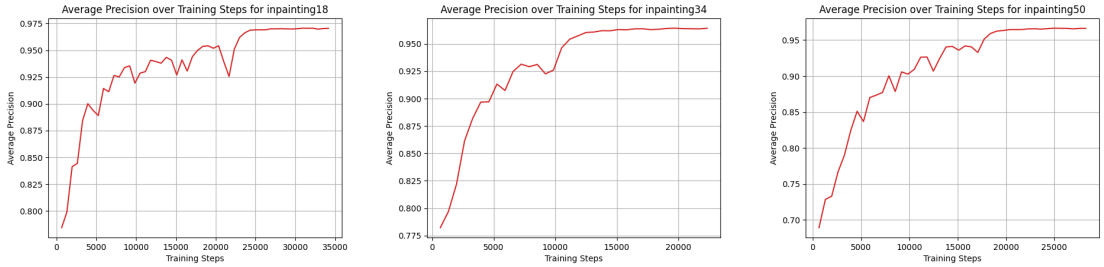
5.2.5. Modelos Basados en el Conjunto de Datos 'insight' de DeepFakeFaces

Los tres siguientes modelos 'insight18', 'insight34' e 'insight50' han sido entrenados a partir de 30.000 imágenes generadas por la herramienta InsightFace [DGXZ19] y 30.000 imágenes reales del conjunto de datos IMDB-WIKI [RTVG15].

Los resultados obtenidos de estos entrenamientos (ver Figura 5.5) mantienen una gran similitud con los resultados obtenidos del entrenamientos de los modelos 'inpainting', tanto en el número de pasos para cada arquitectura como en la puntuación de la precisión. Esto se debe a que la herramienta de InsightFace utiliza la misma técnica de inpainting para realizar los deepfakes, aunque cabe destacar que las trazas de artefactos que deja esta herramienta no coincide con las trazas de artefactos que se encuentran en el conjunto de datos de 'inpainting' como se comentará más adelante, y esto se debe a que la herramienta InsightFace no utiliza Stable Diffusion [RBL+22]. Se comenta que los modelos alcanzan una puntuación de precisión eficaz aunque no tan buena como la de los primeros modelos, explicando así que discriminar este tipo de imágenes sintéticas es más costoso y difícil para



(a) Métrica de precisión (accuracy) durante el entrenamiento.



(b) Métrica de precisión media (average precision) durante el entrenamiento.

(c) Métrica de función de pérdida (loss) durante el entrenamiento.

Figura 5.4: Comparación entre clasificadores con distinta arquitectura ResNet, entrenados con el conjunto de datos 'inpainting' del DeepFakeFace [SHDT23b].

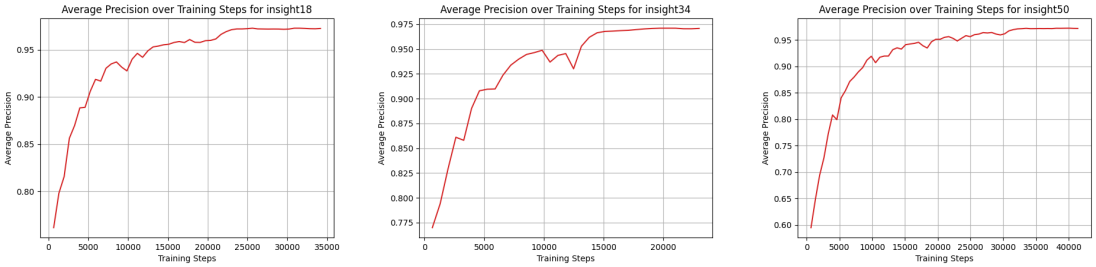
los clasificadores. Se destaca que una vez más es el modelo de arquitectura ResNet34 el más eficiente en el entrenamiento en comparación con los modelos basados en ResNet18 y ResNet50.

5.2.6. Modelo Basado en el Conjunto Total de Datos.

Como se comentará más adelante, se ha comprobado que cada modelo es muy eficaz a la hora de identificar las imágenes con las que ha sido entrenado pero que tiene dificultades a la hora de identificar y clasificar imágenes provenientes de distintos conjuntos de datos. Por ello, se ha entrenado un último clasificador unificando todos los conjuntos de datos anteriores. Se ha decidido utilizar la arquitectura ResNet50 como se recomienda en el artículo [WWZ+20b] debido al gran volumen de imágenes que se va a utilizar en el entrenamiento. Por ello se ha denominado a este modelo como 'every50'.

El conjunto de imágenes sintéticas estaría compuesto por 10.000 imágenes generadas por Stable Diffusion v2.1, 30.000 imágenes generadas por Stable Diffusion v1.4, 30.000

(a) Métrica de precisión (accuracy) durante el entrenamiento.



(b) Métrica de precisión media (average precision) durante el entrenamiento.

(c) Métrica de función de pérdida (loss) durante el entrenamiento.

Figura 5.5: Comparación entre clasificadores con distinta arquitectura ResNet, entrenados con el conjunto de datos 'insight' del DeepFakeFace [SHDT23b].

imágenes generadas por Stable Diffusion v1.4 utilizando la técnica de inpainting y 30.000 imágenes generadas por la herramienta InsightFace.

El conjunto de imágenes reales estaría formado por 10.000 imágenes provenientes de FFHQ y 30.000 imágenes provenientes de IMBD-WIKI.

Se puede observar en la Figura 5.6 que el modelo alcanza una precisión muy eficaz, superior a 0.91, alcanzando más de 80.000 pasos. Este es el modelo que más tiempo ha tardado en alcanzar el earlystopping debido al gran volumen de imágenes que ha usado en su entrenamiento y a la variedad de estas imágenes, que ha aumentado el margen de mejora del propio modelo.

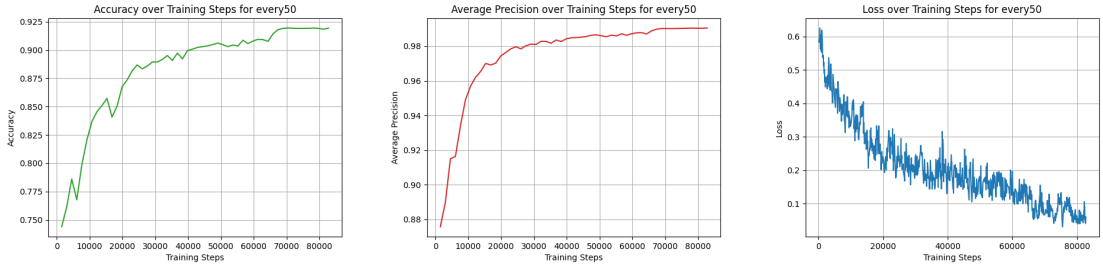


Figura 5.6: Resultados del entrenamiento de 'every50'.

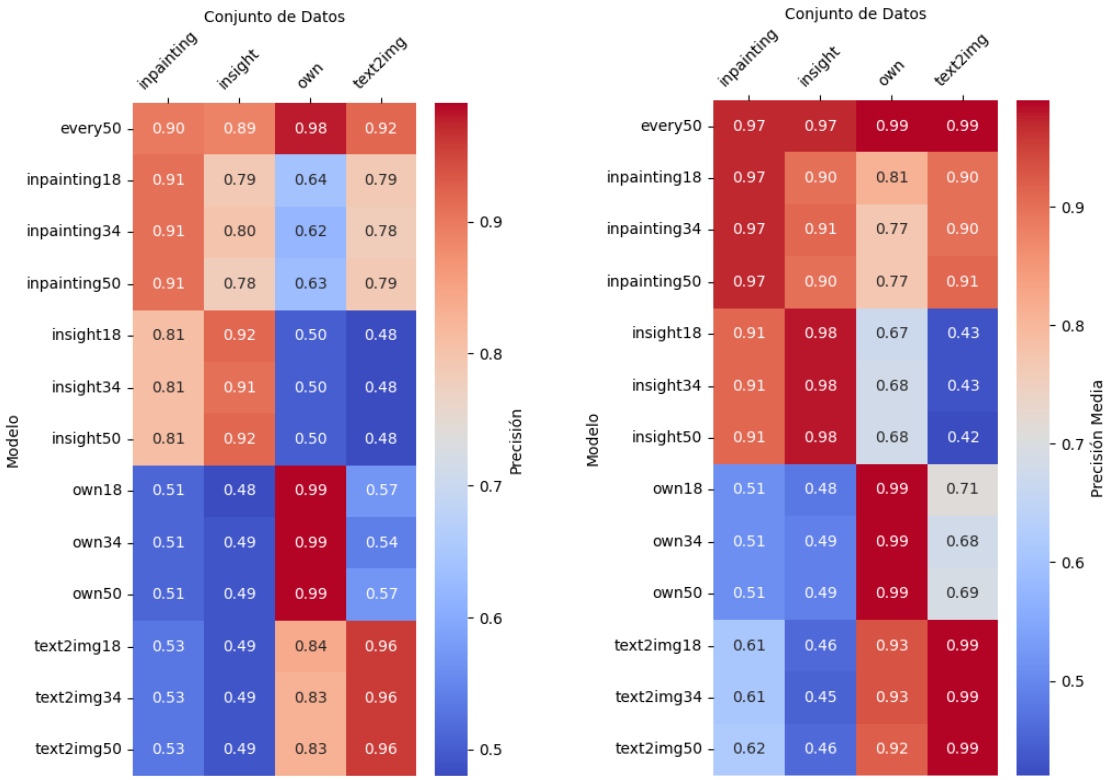
5.3. Comparación de Resultados en Test para los Distintos Modelos Entrenados

A continuación, se muestra la tabla con los resultados de aplicar los distintos clasificadores mencionados en la sección anterior a los conjuntos de datos de pruebas. Cada fila de la Tabla 5.1 representa los resultados logrados por un clasificador en cuanto a precisión (el primer número) y precisión media (el segundo número) para cada conjunto de datos, representados en las distintas columnas.

Tabla 5.1: Resultados de los distintos clasificadores para cada conjunto de datos.

	Conjuntos de Datos de Prueba			
Modelo	**inpainting**	**insight**	**own**	**text2img**
every50	**0.90 / 0.97**	0.89 / **0.97**	**0.98 / 0.99**	**0.92 / 0.99**
inpainting18	**0.91 / 0.97**	0.79 / **0.90**	0.64 / 0.81	0.79 / **0.90**
inpainting34	**0.91 / 0.97**	0.80 / **0.91**	0.62 / 0.77	0.78 / **0.90**
inpainting50	**0.91 / 0.97**	0.78 / **0.90**	0.63 / 0.77	0.79 / **0.91**
insight18	0.81 / **0.91**	**0.92 / 0.98**	0.50 / 0.67	0.48 / 0.43
insight34	0.81 / **0.91**	**0.91 / 0.98**	0.50 / 0.68	0.48 / 0.43
insight50	0.81 / **0.91**	**0.92 / 0.98**	0.50 / 0.68	0.48 / 0.42
own18	0.51 / 0.51	0.48 / 0.48	**0.99 / 0.99**	0.57 / 0.71
own34	0.51 / 0.51	0.49 / 0.49	**0.99 / 0.99**	0.54 / 0.68
own50	0.51 / 0.51	0.49 / 0.49	**0.99 / 0.99**	0.57 / 0.69
text2img18	0.53 / 0.61	0.49 / 0.46	0.84 / **0.93**	**0.96 / 0.99**
text2img34	0.53 / 0.61	0.49 / 0.45	0.83 / **0.93**	**0.96 / 0.99**
text2img50	0.53 / 0.62	0.49 / 0.46	0.83 / **0.92**	**0.96 / 0.99**

También se muestran en la Figura 5.7 dos matrices con mapas de calor aplicados para destacar la puntuación que han logrado los distintos clasificadores.



(a) Matriz de precisión (accuracy). (b) Matriz de precisión media (average precision).

Figura 5.7: Matrices de precisión y precisión media con un mapa de calor aplicado. Cada fila representa un modelo entrenado y cada columna un conjunto de datos de prueba.

Tras analizar los resultados obtenidos, se pueden sacar las siguientes conclusiones:

- La primera es que no se aprecia una diferencia significativa entre los resultados obtenidos por modelos entrenados por un mismo conjunto de datos y con distintas arquitecturas ResNet. Lo cual acota las diferencias entre usar distintas arquitecturas a la eficiencia en los tiempos de entrenamiento y recursos utilizados.
- La segunda conclusión es que cada clasificador logra resultados muy eficaces cuando se trata de clasificar imágenes dentro del conjunto de datos con el que ha entrenado, pero logra resultados mediocres e incluso subóptimos cuando se trata de clasificar imágenes sintéticas procedentes de otros conjuntos de datos.
- La tercera conclusión parte de una continuación de la segunda conclusión, y es que el modelo 'every50' es capaz de lograr una precisión de 0.9 en prácticamente todos los conjuntos de datos, salvo en el conjunto de datos de imágenes generadas por InsightFace [DGXZ19] (se puede observar que es el conjunto de datos más complejo de discriminar si se observa el resto de resultados) donde logra 0.89. Es decir, usando imágenes de conjuntos de datos sintéticos diferentes se puede entrenar un clasificador de carácter generalista capaz de identificar imágenes sintéticas generadas por distintos modelos o a través de diferentes técnicas.

- Se observa una simetría entre las puntuaciones obtenidas por los modelos 'inpainting' e 'insight', y esto se debe a que la herramienta InsightFace [DGXZ19] utiliza la técnica de inpainting para generar sus deepfakes. Dicho de otra manera, ambos conjuntos de imágenes están creados a partir de la misma técnica aunque utilizando distintos modelos de difusión (como se verá más adelante).

- Se observa una asimetría entre las puntuaciones obtenidas por los modelos 'inpainting' y 'text2img', donde los primeros modelos logran clasificar relativamente bien las imágenes del conjunto de datos 'text2img' pero no al revés. Esto se debe a que el conjunto 'inpainting' está generado usando el mismo modelo que el conjunto 'text2img' que es Stable Diffusion [RBL+22] v1.4, pero utilizando la técnica de inpainting, siendo esta una técnica que produce resultados más restrictivos a la hora de ser clasificados. Por ello, los modelos del conjunto 'text2img' no logran clasificar correctamente las imágenes generadas mediante el uso de inpainting mientras que los modelos 'inpainting' logran clasificar correctamente las imágenes del conjunto 'text2img'.

- Se observa otra asimetría entre las puntuaciones obtenidas por los modelos 'own' y 'text2img', ya que ambos modelos están generados con Stable Diffusion [RBL+22]. La asimetría parece ser producida debido a que el conjunto de datos 'own' es muy pequeño y que dispone de menor variedad que el resto de conjuntos, siendo más fácil de clasificar (se observa que otros modelos tienen menos dificultades a la hora de discriminar imágenes en este conjunto de datos).

- Salvando las excepciones comentadas anteriormente, se observa que los modelos que no usan técnicas de inpainting ('own' y 'text2img') tienen dificultades para clasificar los otros conjuntos de datos, y que los modelos que sí usan técnicas de inpainting ('insight' e 'inpainting') tienen dificultades para clasificar los conjuntos de datos generados sin usar esta técnica.

5.4. Análisis de Espectros de Frecuencia para cada Conjunto de Datos

Como se ha comentado anteriormente en la Sección 4.2, se aplicarán una serie de transformaciones al conjunto de imágenes a partir del código propuesto en el proyecto del artículo [RDHF22]. El objetivo es poder identificar diferencias entre las trazas y artefactos que dejan los modelos generativos en comparación con las trazas de colecciones de imágenes reales.

En este caso se compararan dos conjuntos de imágenes reales con los que se han estado realizando entrenamientos y pruebas, que son los conjuntos de imágenes de FFHQ [KLA19] e IMDB-WIKI [RTVG15], y los conjuntos de imágenes sintéticas que consistirían en nuestro propio conjunto 'own' generado usando Stable Diffusion [RBL+22] v2.1, el conjunto de datos 'text2img' generado usando Stable Diffusion v1.4, el conjunto de datos 'inpainting' también generado con Stable Diffusion v1.4 pero utilizando la técnica de inpainting, y el conjunto de datos 'insight' generado utilizando la herramienta InsightFace [DGXZ19].

Se adjuntan las imágenes de los distintos análisis de frecuencia en mayor calidad en el Apéndice B, ya que para apreciar los artefactos encontrados en las trazas de estos análisis

es conveniente mostras las imágenes con una mayor calidad y tamaño.

5.4.1. Análisis de Conjuntos de Datos tras Aplicarles la Transformada de Fourier Discreta.

En este caso se aplica la Transformada Rápida de Fourier (Fast Fourier Transform (FFT)), que consiste en un caso optimizado de la Transformada de Fourier Discreta para el análisis de frecuencia de las imágenes. Se realiza la conversión de imágenes en escala de grises como se ha mencionado anteriormente y se aplica la transformación, obteniendo los resultados expuestos en la Figura 5.8.

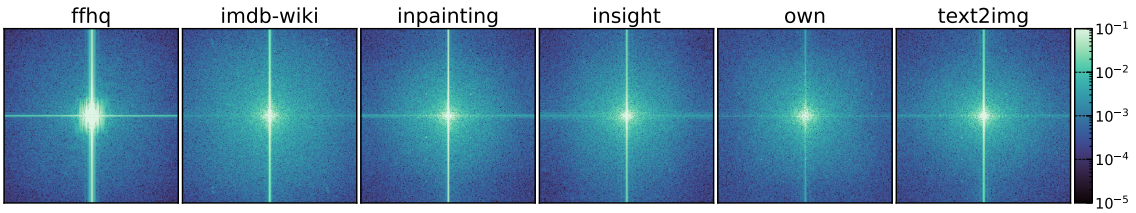


Figura 5.8: Resultados de aplicar la Transformada Rápida de Fourier a los conjuntos de datos usados en el entrenamiento.

- Lo primero que se puede observar es que el conjunto de datos reales de FFHQ muestra una distorsión central en comparación con el resto de conjuntos. Esto se puede deber al formato de compresión del conjunto de datos.
- De la misma manera, se puede distinguir distorsiones cerca de las esquinas del conjunto de datos IMDB-WIKI, y esto se podría explicar por el formato de compresión utilizado en este conjunto de datos.
- Se pueden destacar los artefactos en un patrón de cuadrícula (grid en inglés) que se comentan en los artículos [WWZ+20b, GGB20, RDHF22] que dejan como trazas los modelos de difusión en los conjuntos de datos de 'own', 'text2img' e 'inpainting'. Se destaca que los artefactos son muy similares entre estos dos últimos conjuntos seguramente por estar generados con la misma versión de Stable Diffusion (la versión 1.4), mientras que nuestro conjunto propio tiene unos artefactos menos notorios y esto se podría deber a que está generado con una versión superior de Stable Diffusion (la versión 2.1).
- Es destacable que el conjunto de datos generado con la herramienta InsightFace no genera artefactos visibles como tal para el ojo humano. Aun así, como se ha visto en la Sección 5.2, los clasificadores sí logran discriminar imágenes de este conjunto de datos si son entrenados con éste, por lo que concluímos que, aunque este conjunto de imágenes no deje tras de sí trazas de artefactos en el dominio de Fourier, sí que debe mantener ciertos atributos detectables para los clasificadores.

5.4.2. Análisis de Conjuntos de Datos tras Aplicarles la Transformada de Fourier Discreta con Filtro de Paso Alto.

A continuación, se aplica la Transformada Rápida de Fourier con con un filtro de paso alto incorporado (High Pass - Fast Fourier Transform (HP-FFT)) con el objetivo de

eliminar las frecuencias bajas de la señal, destacando las variaciones rápidas. Se muestran los resultados obtenidos en la Figura 5.9.

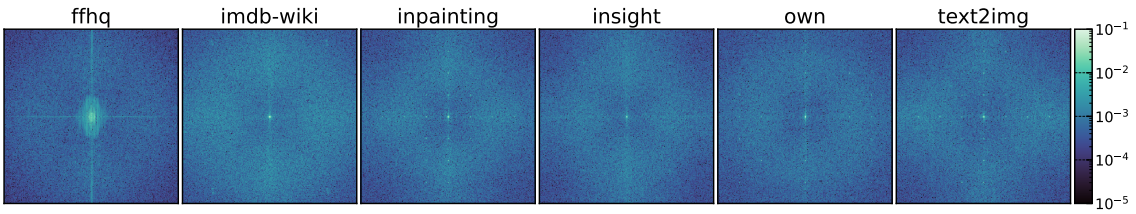


Figura 5.9: Resultados de aplicar la Transformada Rápida de Fourier con Filtro de Paso Alto a los conjuntos de datos usados en el entrenamiento.

- En la Figura 5.9 se puede observar que se destacan aún más las trazas ya observables en la Figura 5.8.
- Se puede percibir de una manera más visual la cuadrícula que surge en las imágenes sintéticas producidas por Stable Diffusion.
- Se aprecia de manera muy sutil por primera vez trazas de esa cuadrícula mencionada previamente en la colección de imágenes de 'insight'.

5.4.3. Análisis de Conjuntos de Datos tras Aplicarles la Transformada del Coseno Discreta.

En este apartado se analizarán los resultados de aplicar la Transformada del Coseno Discreta (DCT) a los conjuntos de imágenes. Como se ha comentado en las Sección 4.2.3, esta función se utiliza para destacar correlaciones espaciales y concentrar componentes de baja frecuencia. Se muestran los resultados obtenidos en la Figura 5.10.

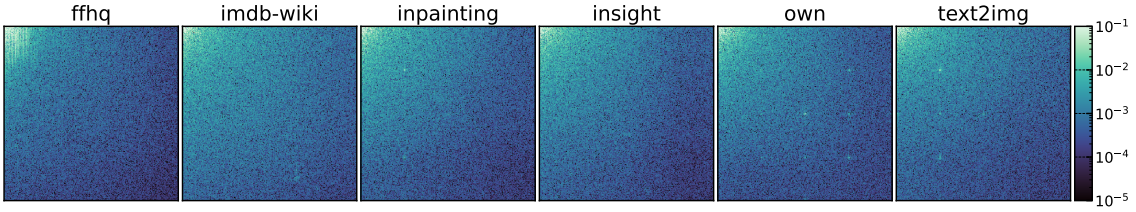


Figura 5.10: Resultados de aplicar la Transformada del Coseno Discreta a los conjuntos de datos usados en el entrenamiento.

- Se pueden observar las trazas en forma de cuadrícula en los conjuntos de datos generados mediante Stable Diffusion, destacando los conjuntos 'own' y 'text2img', mientras que la cuadrícula es menos perceptible en el conjunto 'inpainting'
- El resto de conjuntos parece no mostrar ningún otro tipo de patrón detectable.

5.4.4. Análisis del Espectro de Densidad para cada Conjunto de Datos.

Por último se implantará el cálculo de la densidad del espectro de frecuencia, explicado en la Sección 4.2.4, para cada conjunto de datos.

- Como se puede ver en la Figura 5.11, cada conjunto de datos obtiene un resultado distinto en su análisis del espectro de densidad. Mientras que la densidad entre los conjuntos de datos reales es notoria en todo el espectro, en los conjuntos de datos sintéticos se mantiene de manera similar.
- Las curvas de los conjuntos de datos reales siguen una tendencia paralela, salvando las irregularidades del conjunto de datos de IMDB-WIKI.
- Las curvas de los conjuntos de datos 'own' y 'text2img' muestran una desviación al final de la gráfica, existiendo también en el conjunto de datos 'inpainting' aunque de manera mucho menos pronunciada.
- El conjunto de datos 'insight' no muestra una desviación final como el resto de conjuntos de datos sintéticos pero si muestra desviaciones previas.

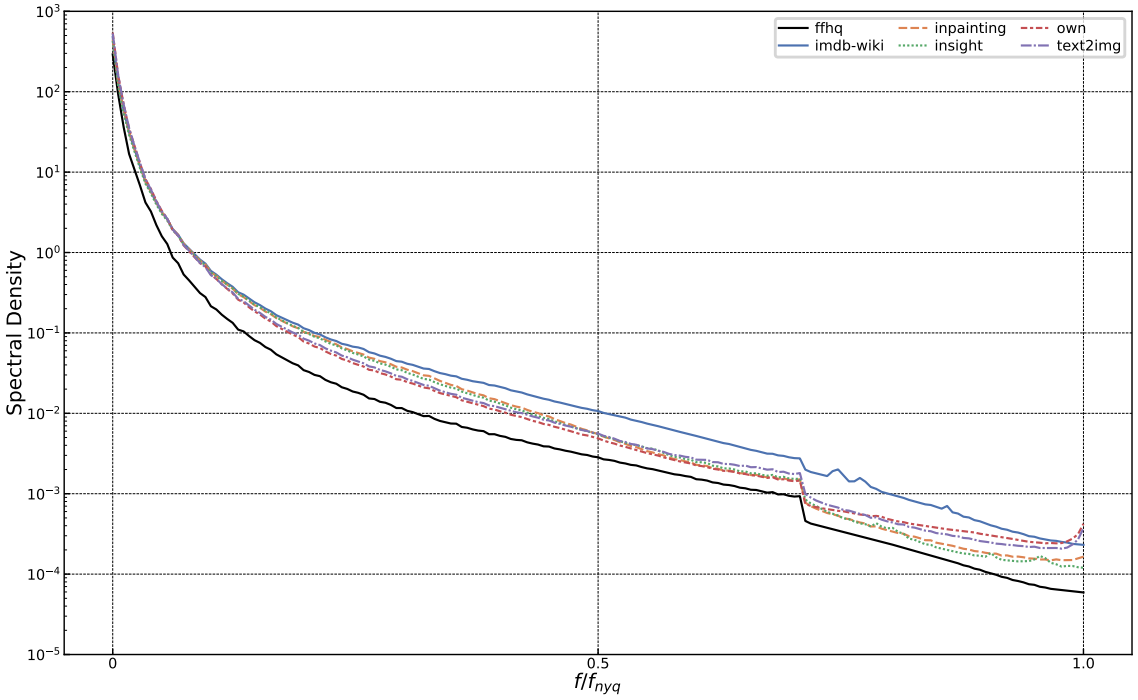


Figura 5.11: Espectro de Densidad de cada conjunto de datos en función de la Frecuencia de Nyquist.

Capítulo 6

Conclusiones y Trabajo Futuro

6.1. Conclusiones

En el presente trabajo se ha realizado una exploración exhaustiva de la evolución y las implicaciones de los modelos de difusión en la generación de contenido sintético. A medida que esta tecnología continúa avanzando a un ritmo vertiginoso, se hace cada vez más esencial comprender sus capacidades y los retos asociados con su regulación y detección.

A continuación, se detallan las conclusiones clave obtenidas a través de nuestra investigación, donde se discuten tanto los avances como los desafíos pendientes en la detección eficaz de imágenes sintéticas y las conclusiones que destacan la necesidad urgente de seguir investigando y desarrollando herramientas que puedan mantenerse a la par con la rápida evolución de estas tecnologías disruptivas.

Una de las primeras conclusiones que se destaca de los resultados obtenidos tras los experimentos realizados en este estudio es que se ha observado que las imágenes sintéticas creadas por modelos de difusión pueden ser identificadas utilizando estrategias como clasificadores binarios basados en arquitecturas ResNet. Estos clasificadores, que han demostrado su eficacia con modelos anteriores como las GANs, según estudios como el de [WWZ+20b], también son efectivos en la detección de imágenes de alta calidad y realismo generadas por modelos de difusión. Aunque estas imágenes son más sofisticadas, los clasificadores aún pueden distinguirlas basándose en características discriminatorias aprendidas. Por lo tanto, se cumple el objetivo relacionado con el estudio del entrenamiento de clasificadores para identificar contenido sintético generado por los modelos de difusión.

Tras analizar los resultados obtenidos también se puede observar que algunos modelos de difusión comparten características y atributos detectables, lo que explica por qué clasificadores entrenados con datos de un modelo específico pueden identificar imágenes de otros modelos de difusión similares. Este hallazgo subraya la posibilidad de desarrollar clasificadores más robustos y versátiles que no solo distingan entre imágenes reales y sintéticas de un solo tipo, sino entre múltiples formas de contenido sintético, aumentando así su aplicabilidad y eficacia en entornos variados.

Sin embargo, aunque se plantea una base inicial mediante el entrenamiento de clasificadores binarios, no se ha logrado desarrollar una herramienta completa para la detección de contenido sintético generado por modelos de difusión, principalmente debido a limitaciones en los recursos disponibles y la complejidad inherente de las tecnologías emergentes. Se plantea la posibilidad de continuar con este desarrollo para producir

herramientas y aplicaciones completas que puedan ser utilizadas por los cuerpos de seguridad en su lucha contra el uso malintencionado del contenido sintético.

Tras las distintas pruebas realizadas durante este proyecto, se demuestra que los modelos de difusión, al igual que sus predecesores, dejan trazas específicas de artefactos en las imágenes que generan, especialmente cuando se examinan en el dominio de Fourier, cumpliendo el objetivo acerca del análisis de imágenes en este dominio. Estas trazas, aunque más sutiles que en modelos como las GANs o los VAEs, dan a entender que estas imágenes sintéticas contienen características y atributos destacables que permiten la identificación y clasificación de éstas. Sin embargo, las técnicas existentes diseñadas para modelos anteriores no son efectivas con los modelos de difusión, lo que indica la necesidad de adaptar o desarrollar nuevas metodologías que puedan destacar y discriminar las imágenes generadas por los emergentes modelos de difusión.

Por último, se establece que esta investigación ha evidenciado una notable escasez de recursos en el campo de la detección de contenido sintético generado por modelos de difusión, destacando especialmente la falta de conjuntos de datos amplios y diversificados. En comparación con los recursos disponibles para modelos generativos más establecidos como las GANs, los modelos de difusión aún no cuentan con el mismo nivel de investigación debido a que son una tecnología todavía emergente.

6.2. Trabajo Futuro

A lo largo de este estudio, se han identificado y abordado numerosas facetas de la tecnología de modelos de difusión, revelando tanto su potencial como sus desafíos inherentes. Si bien se ha avanzado significativamente en la investigación de estos modelos, la rápida evolución y los cambios constantes de estos modelos indican que hay mucho más por investigar y muchas oportunidades para innovaciones futuras.

En esta sección, se delinearán posibles direcciones para investigaciones futuras con el objetivo de implementar y mejorar las aplicaciones prácticas de esta tecnología:

- Dada la rápida evolución de los modelos de difusión y su creciente aplicación en la generación de contenido sintético, se hace imperativo para la comunidad científica disponer de conjuntos de datos amplios y variados que emulen los ya existentes para otros modelos generativos como las GANs. Estos conjuntos de datos son esenciales para el desarrollo y la mejora de herramientas que permitan detectar y diferenciar eficazmente entre contenido sintético y real. Actualmente, la disponibilidad de estos datos es limitada debido a la novedad y naturaleza emergente de la tecnología de modelos de difusión. Por tanto, una línea de trabajo futuro imperativa sería la creación de múltiples conjuntos de datos robustos que faciliten la investigación y el desarrollo de nuevas metodologías de detección.

- Se propone continuar y expandir la línea de investigación actual para incluir clasificadores que se entrenen utilizando datos sintéticos generados no solo por modelos de difusión, sino también por otros modelos generativos como GANs, VAEs, Transformadores o CNNs. Algunos estudios preliminares, como el mencionado en el artículo [GGB23], ya han comenzado a explorar clasificadores anidados que pueden determinar el tipo de modelo generativo responsable de la creación de una imagen específica. El desarrollo de estos sistemas clasificatorios avanzados es de gran interés

para la comunidad científica ya que con ellos se propone mejorar la precisión y la eficacia en la identificación del origen de las imágenes sintéticas.

- Otro enfoque de investigación prometedor consiste en el desarrollo de técnicas alternativas de detección que complementen los métodos existentes para identificar imágenes sintéticas producidas por modelos de difusión. Por ejemplo, el artículo [WBZ+23] sugiere la aplicación de técnicas de reconstrucción usando técnicas de difusión y reconstrucción gaussiana, que emulan los procesos subyacentes en los modelos de difusión. Explorar y perfeccionar técnicas como ésta podría proporcionar nuevas vías para el análisis y la verificación de la autenticidad del contenido visual.

- Finalmente, sería de gran beneficio desarrollar una aplicación práctica que integre estos detectores ya entrenados, con el objetivo de facilitar su uso de manera sencilla y accesible. Tal herramienta podría ser de un valor altísimo para las fuerzas de seguridad y otros organismos encargados de la aplicación de la ley, permitiéndoles combatir eficazmente el uso malicioso de deepfakes y contribuir significativamente en la lucha contra la proliferación de contenido sexual ilícito que involucra a menores.

Capítulo 7

Introduction

7.1. Motivation

Over the past few decades, we have witnessed a digital revolution that has dramatically transformed the technological world, particularly in the realm of image creation and manipulation. The development of generative models such as GANs, VAEs, and more recently, diffusion models, has marked a significant advance in how we can generate high-quality synthetic images that are nearly indistinguishable from real ones.

In the early days of artificial intelligence applied to image generation, hardware and software limitations restricted the quality and complexity of the generated images. However, with the evolution of digital technology, these models have greatly improved, resulting in images that not only deceive the human eye but also present significant challenges for detection using conventional methods.

The shift towards more advanced models like diffusion models has further complicated the task of differentiating between real and synthetic images. Although initially designed to improve the quality of generated images, their ability to produce hyper-realistic results has opened new doors for both positive and potentially malicious applications, such as the creation of misinformation or fake content. The capability of diffusion models to generate images indistinguishable from real ones has led to an alarming increase in the production of inappropriate content, including pornographic material involving representations of minors. Such content, although synthetic, is deeply problematic and remains illegal in many jurisdictions around the world. The ease and accessibility with which these models can generate hyper-realistic images have resulted in an exponential increase in child sexual abuse material in the digital realm, which is deeply alarming and requires urgent action.

On the other hand, while detection methods developed for earlier generative models have proven to be relatively effective, adapting well to the specific characteristics of the images they generate, these techniques find significant limitations when applied to diffusion models, which employ different and more sophisticated mechanisms for image creation. Given this scenario, it is imperative not only to adapt existing detection methods but also to develop new techniques that can effectively identify the complexities of images generated by the latest models.

This work aims to study in-depth the current situation of synthetic image generation, especially those produced by diffusion models, which represent the most advanced frontier in generative technologies. Furthermore, the current state of synthetic content detection

techniques will be evaluated to identify their limitations and areas for improvement. From this evaluation, the goal is to develop, test, and propose solutions that enhance the ability to identify and discriminate between real and synthetic images generated by diffusion models. The purpose of this effort is twofold: on one hand, to improve digital security and, on the other, to provide reliable tools for forensic image analysis in a context dominated by artificial intelligence.

7.2. Context

This Master's Thesis is part of a research project entitled "Novel Strategies to Fight Child Sexual Exploitation and Human Trafficking Crimes and Protect their Victims – HEROES,.[a]pproved by the European Commission within the Horizon 2020 Framework Programme (H2020-SU-SEC-2020 call) under grant agreement number 101021801. The project coordinator is the GASS Group from the Complutense University of Madrid (Group of Analysis, Security and Systems, https://gass.ucm.es, group 910623 in the catalog of research groups recognized by the UCM).

In addition to the Complutense University of Madrid, HEROES includes 24 entities located in 17 countries: 11 from EU countries (Austria, Belgium, Bulgaria, France, Greece, Ireland, Latvia, Lithuania, Portugal, Spain, the United Kingdom), 1 associated country (Switzerland) and 5 third countries (Bangladesh, Brazil, Colombia, Peru, Uruguay). These entities are: University of Kent (United Kingdom), The Free University of Brussels (Belgium), The French National Research Institute for Digital Science and Technology – INRIA (France), Center for Security Studies – KEMEA (Greece), International Centre for Migration Policy Development – ICMPD (Austria), International Center for Missing and Exploited Children – ICMEC (Switzerland), IDENER Research & Development Economic Interest Grouping (Spain), Athena Research Center – ARC (Greece), Trilateral Research and Consulting (United Kingdom), Centre for Women and Children Studies – CWCS (Bangladesh), Center Against Human Trafficking and Exploitation – KOPZI (Lithuania), Portuguese Association for Victim Support – APAV (Portugal), Fundación Renacer (Colombia), The Greek Council for Refugees – GCR (Greece), Brazilian Association for the Defense of Children and Youth – ASBRAD (Brazil), Hellenic Police (Greece), Latvia National Police (Latvia), General Directorate for the Fight against Organized Crime (Bulgaria), General Directorate of the Police – DGP (Spain), Federal Police (Brazil), Federal Highway Police (Brazil), Secretariat for Strategic Intelligence of the State – Presidency of the Oriental Republic of Uruguay (Uruguay).

7.3. Research Object

The detection of synthetic images generated using diffusion models poses unique challenges compared to previous generative models such as GANs. Although diffusion models produce high-quality, realistic images, recent research has shown that, like their predecessors, they also leave detectable traces and artifacts that can be exploited to differentiate between real and synthetic images. These traces may not be perceptible at first glance, but they are identifiable through advanced image analysis techniques.

This work focuses on studying these specific residual traces left by diffusion models, exploring how convolutional neural networks with architectures such as ResNet can be

trained to classify images in terms of their authenticity. These networks are capable of learning and recognizing complex patterns in visual data that are often invisible to humans, offering a powerful tool in the fight against visual misinformation and deepfakes.

Furthermore, the analysis of images in the Fourier domain will be addressed, a technique that allows the frequencies of images to be studied to detect subtle variations that are not evident in the temporal space. This approach is particularly useful for identifying specific signatures that diffusion models leave on synthetic images, such as anomalies in textures or unusual patterns in the objects contained in the images, which are indicative of synthetic content.

The combination of these methodological approaches aims to develop a robust detection system that can be used not only to identify images generated by diffusion models but also to provide a solid foundation for the development of future detection tools as these technologies evolve. This study will contribute significantly to the field of digital security, providing insights and resources to effectively combat the malicious use of synthetic images.

7.4. Work Plan

The development of this work is proposed in four phases:

1. **Research**: The research phase will focus on the first four months, where the project will be dedicated to understanding and learning about the context of generative models. During this period, a thorough study of the existing literature on generative models such as GANs, VAEs, and Transformers will be conducted, along with a specific and extensive analysis of diffusion models. Regular meetings will be held to discuss progress, resolve doubts, and share useful resources like Google Scholar for searching relevant articles. This phase will also include exploring data and previous techniques in the detection of synthetic images, with the aim of highlighting the scarcity of specific information about diffusion models, which will outline the path toward experimentation and the development of new tools.

2. **Development**: After consolidating the theoretical foundation, the project will move into the development phase, with the goal of focusing efforts on generating a proprietary dataset using recent latent diffusion tools such as Stable Diffusion [RBL+22], given the limited access to specific datasets of images generated by diffusion models. Additionally, neural network architectures such as ResNet will be prepared and adjusted to train with these new datasets. During this stage, transformations in the Fourier domain will be applied to closely observe the traces and residual artifacts generated by the diffusion models, techniques used to identify distinctive features in the synthetic images produced by previous generative models.

3. **Experimentation**: The third phase will involve the practical implementation of binary classifiers to test their effectiveness in discriminating between real and synthetic images. Experiments will be conducted with different ResNet configurations, adjusting parameters and evaluating their performance with multiple datasets to ensure that the models can generalize and detect more than one type of synthetic image. Comparisons of the results obtained will be made, and adjustments will be made to the architectures and parameters of the models with the aim of optimizing these models.

4. **Documentation**: In the final phase, time will be dedicated to documenting each stage of the process, from the initial research phase to the final tests of the classifiers, with the goal of creating a comprehensive document for the Master's Thesis, serving as a collection of all the work done and as a valuable resource for future research in the field of synthetic image detection.

7.5. Structure of the Work

The rest of the work is organized into six chapters and two appendices with the following structure: Chapter 2 introduces some concepts that are fundamental to understanding generative models, exploring different types like GANs, CNNs, VAEs, ARs, Transformers, and particularly, diffusion models. The processes of diffusion and reversal are discussed, and the different types of diffusion models that exist are described, establishing the necessary theoretical basis for the following chapters.

Chapter 3 addresses the current state of synthetic image detection. The chapter begins with a review of how detection techniques work in previous generative models and then focuses on the detection of images generated by diffusion models. The applicability of existing techniques for previous models and the necessary adaptation for diffusion models are discussed. The available datasets are also analyzed, highlighting the scarcity of data for diffusion models and the analysis of images in the Fourier domain.

Chapter 4 presents the methodology used in this project, describing the binary classifiers, ResNet architectures, and the use of different transformations for frequency analysis. It explains how the datasets used were prepared, including the generation of a proprietary set, and the planning of the project to address the established research objectives.

Chapter 5 describes the experiments conducted to evaluate the effectiveness of the algorithms proposed in Chapter 4. This chapter covers everything from setting up model training using various ResNet architectures and data augmentation techniques to evaluating these models using different datasets. A detailed analysis of the results is included, comparing different configurations and applying frequency analysis to check the traces left by diffusion models in synthetic images.

Chapter 6 presents the conclusions derived from the research, summarizing the main findings and discussing their implications. Future lines of research are also outlined, proposing how these studies can be expanded and adapted to continue improving the detection of synthetic images in an ever-evolving field.

Chapters 7 and 8 are the English translations of the Introduction and Conclusions.

Appendix A demonstrates the operation of a script proposed for generating a dataset of synthetic images produced by Stable Diffusion [RBL+22].

Appendix B collects results generated in this research with the goal of providing higher quality and larger size images to properly appreciate the traces of artifacts produced by diffusion models in the Fourier domain.

Capítulo 8

Conclusions and Future Work

8.1. Conclusions

In this work, an exhaustive exploration of the evolution and implications of diffusion models in the generation of synthetic content has been conducted. As this technology continues to advance at a rapid pace, it becomes increasingly essential to understand its capabilities and the challenges associated with its regulation and detection.

The following are the key conclusions obtained through our research, discussing both the advances and the pending challenges in the effective detection of synthetic images. These conclusions underscore the urgent need to continue researching and developing tools that can keep pace with the rapid evolution of these disruptive technologies.

One of the initial conclusions from the results obtained after the experiments conducted in this study is that synthetic images created by diffusion models can be identified using strategies such as binary classifiers based on ResNet architectures. These classifiers, which have proven effective with earlier models such as GANs, according to studies like [WWZ+20b], are also effective in detecting high-quality and realistic images generated by diffusion models. Although these images are more sophisticated, the classifiers can still distinguish them based on learned discriminatory features. Thus, the objective related to studying the training of classifiers to identify synthetic content generated by diffusion models has been met.

Upon analyzing the obtained results, it is also observed that some diffusion models share detectable characteristics and attributes, explaining why classifiers trained with data from a specific model can identify images from other similar diffusion models. This finding underscores the possibility of developing more robust and versatile classifiers that not only distinguish between real and synthetic images of a single type but across multiple forms of synthetic content, thus increasing their applicability and effectiveness in varied environments.

However, although an initial basis is established through the training of binary classifiers, a complete tool for detecting synthetic content generated by diffusion models has not been developed, mainly due to limitations in available resources and the inherent complexity of emerging technologies. There is the possibility to continue this development to produce comprehensive tools and applications that can be used by law enforcement in their fight against the malicious use of synthetic content.

Throughout the various tests conducted during this project, it has been demonstrated that diffusion models, like their predecessors, leave specific traces of artifacts in the images they generate, especially when examined in the Fourier domain, fulfilling the objective regarding the analysis of images in this domain. These traces, although more subtle than in models such as GANs or VAEs, suggest that these synthetic images contain notable characteristics and attributes that allow for their identification and classification. However, existing techniques designed for previous models are not effective with diffusion models, indicating the need to adapt or develop new methodologies that can highlight and discriminate the images generated by emerging diffusion models.

Finally, it is established that this research has evidenced a notable scarcity of resources in the field of detecting synthetic content generated by diffusion models, particularly highlighting the lack of comprehensive and diversified datasets. Compared to the resources available for more established generative models such as GANs, diffusion models still do not have the same level of research due to their still emerging technology status.

8.2. Future Work

Throughout this study, numerous facets of diffusion model technology have been identified and addressed, revealing both its potential and inherent challenges. While significant progress has been made in researching these models, the rapid evolution and constant changes of these models suggest that there is much more to explore and many opportunities for future innovations.

In this section, possible directions for future research will be outlined with the aim of implementing and improving the practical applications of this technology:

- Given the rapid evolution of diffusion models and their increasing application in the generation of synthetic content, it is imperative for the scientific community to have access to broad and varied datasets that emulate those already existing for other generative models such as GANs. These datasets are essential for the development and improvement of tools that allow for effective detection and differentiation between synthetic and real content. Currently, the availability of these data is limited due to the novelty and emerging nature of diffusion model technology. Therefore, an imperative future line of work would be the creation of multiple robust datasets that facilitate research and the development of new detection methodologies.

- It is proposed to continue and expand the current line of research to include classifiers that are trained using synthetic data generated not only by diffusion models but also by other generative models such as GANs, VAEs, Transformers or CNNs. Preliminary studies, such as the one proposed by Guarnera et al. [GGB23], have already begun to explore nested classifiers that can determine the type of generative model responsible for the creation of a specific image. The development of these advanced classification systems is of great interest to the scientific community as it proposes to improve the precision and efficacy in identifying the origin of synthetic images.

- Another promising research approach involves the development of alternative detection techniques that complement existing methods for identifying synthetic images produced by diffusion models. For example, the article [WBZ+23] by Wang et al. suggests the application of reconstruction techniques using diffusion and

Gaussian reconstruction techniques, which emulate the underlying processes in diffusion models. Exploring and refining techniques like this could provide new avenues for the analysis and verification of the authenticity of visual content.

- Finally, it would be of great benefit to develop a practical application that integrates these already trained detectors, with the aim of facilitating their use in a simple and accessible manner. Such a tool could be of immense value to security forces and other law enforcement agencies, enabling them to effectively combat the malicious use of deepfakes and significantly contribute to the fight against the proliferation of illicit sexual content involving minors.

Bibliografía

[AI24] Stability AI. Stability ai - home page, 2024.

[AKA23] Agil Aghasanli, Dmitry Kangin, and Plamen Angelov. Interpretable-through-prototypes deepfake detection for diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 467–474, 2023.

[Ama24] Amazon Web Services. Amazon web services (aws) - cloud computing services, 2024.

[Ba19] Hung Ba. Improving detection of credit card fraudulent transactions using generative adversarial networks. *arXiv preprint arXiv:1907.03355*, 2019.

[Ben19] Irina Kofman JE Tester JLElliott Joshua Metherd Julia Elliott Mozaic Phil Culliton Sohier Dane Woo Kim Benpflaum, Brian G. Deepfake detection challenge, 2019.

[BRL+23] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[CCK+18] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[CCP+23] 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.

[CCZ+23] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[CGdMG24] Daniel Cabañas González, Universidad Complutense de Madrid, and Grupo GASS. Análisis de Técnicas de Detección de Imágenes Sintéticas, 2024. Código fuente disponible en: https://github.com/cabannas/analysis-of-synthetic-image-detection-techniques

[CHIS23] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[CTG+24] Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[Cue23] Pedro Cuenca. LSUN Bedrooms Dataset on Hugging Face, 2023.

[CWG+21] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021.

[DBP+] B Dolhansky, J Bitton, B Pflaum, J Lu, R Howes, M Wang, and CC Ferrer. The deepfake detection challenge (dfdc) dataset. arxiv 2020. *arXiv preprint arXiv:2006.07397*.

[DDS+09] 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. Ieee, 2009.

[DGXZ19] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[DYH+21] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021.

[FCJ+23] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22466–22477, 2023.

[FDSP+19] Ugo Fiore, Alfredo De Santis, Francesca Perla, Paolo Zanetti, and Francesco Palmieri. Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences*, 479:448–455, 2019.

[FFDR+21] Li Fei-Fei, Jia Deng, Olga Russakovsky, Alex Berg, and Kai Li. ImageNet, 2021.

[G.21] Ajaykumar G. LSUN Church Dataset on Kaggle, 2021.

[GCM+21] 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 *2021 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2021.

[GGB20] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 666–667, 2020.

[GGB23] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Level up the deepfake detection: a method to effectively discriminate images generated by gan architectures and diffusion models. *arXiv preprint arXiv:2303.00608*, 2023.

[GLZ+23] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10021–10030, 2023.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[GYR+23] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[JLW+20] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2889–2898, 2020.

[JN23] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023.

[KALL17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[KLA19] 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*, pages 4401–4410, 2019.

[KM18] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018.

[KM19] Pavel Korshunov and Sébastien Marcel. Vulnerability assessment and detection of deepfake videos. In *2019 International Conference on Biometrics (ICB)*, pages 1–6. IEEE, 2019.

[LAI24] LAION. Laion ai, 2024.

[LCB10] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[LCH+06] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

[LDK23] Peter Lorenz, Ricard L Durall, and Janis Keuper. Detecting images generated by deep diffusion models using their local intrinsic dimensionality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 448–459, 2023.

[LDR+22] 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*, pages 11461–11471, 2022.

[LGS+23] Zhenghao Lin, Yeyun Gong, Yelong Shen, Tong Wu, Zhihao Fan, Chen Lin, Nan Duan, and Weizhu Chen. Text generation with diffusion language models: A pre-training approach with continuous paragraph denoise. In *International Conference on Machine Learning*, pages 21051–21064. PMLR, 2023.

[LLWT15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[LLWT21] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. CelebA Dataset, 2021.

[LMB+14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[LPR+21] Tsung-Yi Lin, Genevieve Patterson, Matteo R. Ronchi, Yin Cui, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. COCO Dataset, 2021.

[LTG+22] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022.

[Luo22] Calvin Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022.

[LYS+20] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216, 2020.

[ND21] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

[NKH19] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7184–7193, 2019.

[PGH+16] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. *Advances in neural information processing systems*, 29, 2016.

[PTD+22] Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, pages 117–126. Springer, 2022.

[PVZJ12a] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. The Oxford-IIIT Pet Dataset, 2012.

[PVZJ12b] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

[RBL+22] 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*, pages 10684–10695, 2022.

[RCV+18] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. arxiv 2018. *arXiv preprint arXiv:1803.09179*, 2018.

[RCV+19] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.

[RDHF22] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. Towards the detection of diffusion model deepfakes. *arXiv preprint arXiv:2210.14571*, 2022.

[RDN+22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arxiv 2022. *arXiv preprint arXiv:2204.06125*, 2022.

[RDS+15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[RKH+21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[RLJ+23] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[RNMS22] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H Sung. Deepfake detection: A systematic literature review. *IEEE access*, 10:25494–25513, 2022.

[RTVG15] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 10–15, 2015.

[SBV+22] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[SC21] Divya Saxena and Jiannong Cao. Generative adversarial networks (gans) challenges, solutions, and future directions. *ACM Computing Surveys (CSUR)*, 54(3):1–42, 2021.

[SCS+22] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[SDWMG15] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[SHDT23a] Haixu Song, Shiyu Huang, Yinpeng Dong, and Wei-Wei Tu. DeepFakeFace Dataset, 2023.

[SHDT23b] Haixu Song, Shiyu Huang, Yinpeng Dong, and Wei-Wei Tu. Robustness and generalizability of deepfake detection: A study with diffusion models. *arXiv preprint arXiv:2309.02218*, 2023.

[SME20] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[Sta24] Stability AI and Hugging Face. Stable diffusion 2.1 on hugging face, 2024.

[SVB+21] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[The23] The Computer Vision Foundation. International conference on computer vision (iccv) 2023, 2023.

[TN23] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023.

[TZS+16] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

[Ult23] Ultralytics. MNIST Dataset Documentation, 2023.

[VdOKE+16] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.

[VK22] Arash Vahdat and Karsten Kreis. Improving diffusion models as an alternative to gans, part 1, 2022.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[WBZ+23] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023.

[Wes19] Mika Westerlund. The emergence of deepfake technology: A review. *Technology innovation management review*, 9(11), 2019.

[WWZ+20a] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN Detection Github Project, 2020.

[WWZ+20b] 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*, pages 8695–8704, 2020.

[YLL19] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019.

[YLY+18] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

[YSZ+15] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[YXK+22] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

[YZS+23] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

[ZCC+20] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2382–2390, 2020.

[ZCY+24a] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems*, 36, 2024.

[ZCY+24b] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. GenImage Dataset, 2024.

[ZJZ+23] Guanhua Zhang, Jiabao Ji, Yang Zhang, Mo Yu, Tommi S Jaakkola, and Shiyu Chang. Towards coherent image inpainting using denoising diffusion implicit models. 2023.

[ZKC19] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2019.

[ZPIE17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[ZTC22] Qinsheng Zhang, Molei Tao, and Yongxin Chen. gddim: Generalized denoising diffusion implicit models. *arXiv preprint arXiv:2206.05564*, 2022.

[ZWS+22] Xian Zhang, Xin Wang, Canghong Shi, Zhe Yan, Xiaojie Li, Bin Kong, Siwei Lyu, Bin Zhu, Jiancheng Lv, Youbing Yin, et al. De-gan: Domain embedded gan for high quality face image inpainting. *Pattern Recognition*, 124:108415, 2022.

[ZZCH18] Junhai Zhai, Sufang Zhang, Junfen Chen, and Qiang He. Autoencoder and its various variants. In *2018 IEEE international conference on systems, man, and cybernetics (SMC)*, pages 415–419. IEEE, 2018.

Apéndices

Apéndice A

Generación Automatizada de Conjunto de Datos

En este apéndice se muestra en qué consiste y cómo funciona el script `dataset_generator.py`, mencionado en la Sección 4.4.1.

A.1. Descripción General del Script

El script `dataset_generator.py` está diseñado para automatizar la creación de un conjunto de datos compuesto por imágenes sintéticas, utilizando la tecnología de generación de imágenes mediante el modelo Stable Diffusion [RBL+22] v2.1 desarrollado por Stability AI [AI24] a través de la API de HuggingFace [Sta24]. Este script genera un número de imágenes definido por el parámetro `num_imgs`, estando cada una de estas imágenes configurada con atributos visuales específicos basados en parámetros predefinidos y seleccionados aleatoriamente que se explicarán a continuación.

```
import torch
from diffusers import StableDiffusionPipeline, DPMSolverMultistepScheduler
from tqdm import tqdm
import random
import os
```

A.2. Detalles de los Parámetros y Opciones de Atributos

El script `dataset_generator.py` utiliza una serie de parámetros configurables y opciones de atributos para personalizar la generación de imágenes. A continuación, se detallan los principales componentes y su función:

A.2.1. Parámetros Configurables

- **Número de imágenes**: Define el número total de imágenes a generar, especificado por el parámetro `num_imgs`. Este valor está configurado inicialmente en `10000`.
- **Número de inicio para renombrar**: Determina el número inicial para el nombramiento secuencial de las imágenes generadas, controlado por el parámetro

`num_rename` y configurado inicialmente a `0`. El propósito de este parámetro es el de no sobreescribir imágenes si se está generando el conjunto de datos de manera progresiva. Es decir, si se generasen 10.000 imágenes iniciales y se quisiera seguir generando imágenes se debería inicializar este parámetro al número total de imágenes que se tuviesen en ese momento, que sería el número `10000`.

```
num_images = 10000 # Total number of images you want to generate
num_rename = 0 # Number in which naming starts
```

A.2.2. Opciones de Atributos

Cada imagen se genera con atributos personalizables seleccionados aleatoriamente basados en pesos definidos. Los atributos y sus opciones son:

- **Raza**: Puede ser `asian`, `african`, `hindu`, o `caucasian`, con una distribución personalizable de pesos en la que se debe indicar el porcentaje para cada raza.
- **Gafas**: Las opciones son `glasses` y `no glasses`, con pesos personalizables establecidos incialmente en 30 % y 70 % respectivamente.
- **Color de ojos**: Incluye `blue eyes`, `green eyes` y `brown eyes`, con pesos personalizables establecidos inicialmente en 10 %, 10 % y 80 %.
- **Longitud del cabello**: Las opciones son `curly hair`, `short hair`, y `long hair`, con pesos personalizables inicialmente establecidos en 15 %, 45 %, y 40 %.
- **Pecas**: Puede ser `freckles` o `no freckles`, con pesos personalizables establecidos inicialmente en 20 % y 80 %.
- **Expresión facial**: Incluye `smiling` y `serious`, distribuidas equitativamente al 50 %, también personalizables.
- **Edad y género**: Contiene opciones como `baby boy`, `baby girl`, `young boy`, etc., con distintas distribuciones de peso también personalizables.
- **Fondo**: Las opciones de fondo incluyen `office`, `urban street`, y `park`, con distribuciones de peso equitativas y personalizables.

```
# Define atributes and weights

race_options = [('asian', 20), ('african', 20), ('hindu', 20), ('caucasian', 40)]
glasses_options = [('glasses', 30), ('no glasses', 70)]
eye_color_options = [('blue eyes', 10), ('green eyes', 10), ('brown eyes', 80)]
hair_length_options = [('curly hair', 15), ('short hair', 45), ('long hair', 40)]
freckles_options = [('freckles', 20), ('no freckles', 80)]
facial_expression_options = [('smiling', 50), ('serious', 50)]
age_n_gender_options = [('baby boy', 10), ('baby girl', 10),
                        ('young boy', 10), ('young girl', 10),
                        ('teenager boy', 10), ('teenager girl', 10),
                        ('man', 15), ('woman', 15),
                        ('old man', 5), ('old woman', 5)]
background_options = [('office', 33), ('urban street', 33), ('park', 34)]
```

Cada combinación de atributos resulta en un *prompt* dinámico que se utiliza para instruir al modelo de generación de imágenes, garantizando una diversidad en las imágenes generadas que refleja la variabilidad real de las características humanas en distintos contextos ambientales.

```
def weighted_choice(options):
    choices, weights = zip(*options)
    return random.choices(choices, weights=weights, k=1)[0]

def create_prompt():
    race = weighted_choice(race_options)
    glasses = weighted_choice(glasses_options)
    eye_color = weighted_choice(eye_color_options)
    hair_length = weighted_choice(hair_length_options)
    freckles = weighted_choice(freckles_options)
    facial_expression = weighted_choice(facial_expression_options)
    age_n_gender = weighted_choice(age_n_gender_options)
    background = weighted_choice(background_options)

    # Modify prompt at will
    prompt = f"a␣centered␣portrait␣of␣a␣{facial_expression}␣{race}
␣␣␣␣{age_n_gender}␣with␣{eye_color},␣{hair_length},␣{glasses}␣and
␣␣␣␣{freckles},␣in␣a␣{background}"
    return prompt
```

A.3. Generación de Imágenes

El script utiliza el modelo de difusión estable Stable Diffusion de Stability AI, específicamente la versión 2.1, para la generación de imágenes. Este modelo es accesible a través de la biblioteca 'diffusers' de Hugging Face y se optimiza para operar con precisión de punto flotante de 16 bits (`torch.float16`) en una GPU, lo que mejora tanto el rendimiento como la eficiencia en el uso de la memoria.

```
model_id = "stabilityai/stable-diffusion-2-1"
pipe = StableDiffusionPipeline.from_pretrained(model_id, torch_dtype=torch.float16)
pipe.scheduler = DPMSolverMultistepScheduler.from_config(pipe.scheduler.config)
pipe = pipe.to("cuda")

# Create output directory
output_dir = "stable_diffusion/fake_dataset"
os.makedirs(output_dir, exist_ok=True)
```

El código se ejecuta en un bucle 'for' tantas veces como el parámetro establecido `num_imgs`:

```
# Generate and save images
for i in tqdm(range(num_images), desc="Generating␣images"):
    # Randomize the seed for each image
    seed = torch.seed()
    prompt = create_prompt()

    # Generate image
    generator = torch.Generator(device='cuda').manual_seed(seed)
    image = pipe(prompt, height=512, width=512, generator=generator).images[0]

    # Save image
```

```
    filename = f"{output_dir}/fi_{i+num_rename}.png"
    image.save(filename)
    print(f"Saved␣{filename}")

print("Dataset␣generation␣complete!")
```

Cada imagen generada se guarda en el directorio especificado, siguiendo un esquema de nomenclatura que comienza desde un número definido por el usuario con el parámetro establecido `num_rename`.

Con este script se tiene el objetivo de lograr una generación de un conjunto de datos diverso y de gran magnitud, siendo este script una herramienta muy interesante para la comunidad cientifica en el campo del aprendizaje profundo y la visión computacional.

Apéndice B

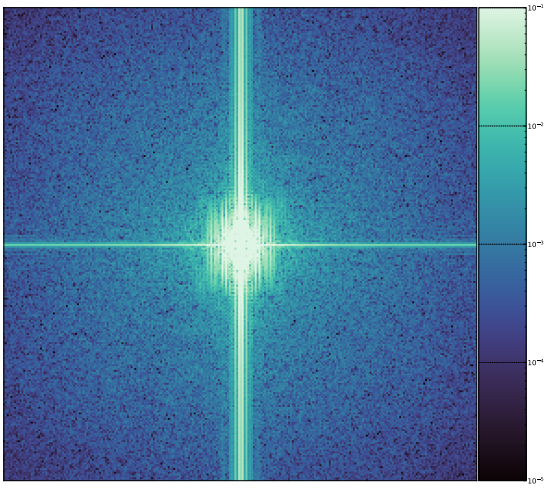
Resultados del Análisis de Espectros de Frecuencia para cada Conjunto de Datos

En este apéndice se adjuntan las imágenes comentadas en la Sección 5.4 en una mayor calidad y tamaño, con el objetivo de poder apreciar con mayor calidad las trazas de artefactos que dejan los modelos generativos a la hora de producir imágenes sintéticas.

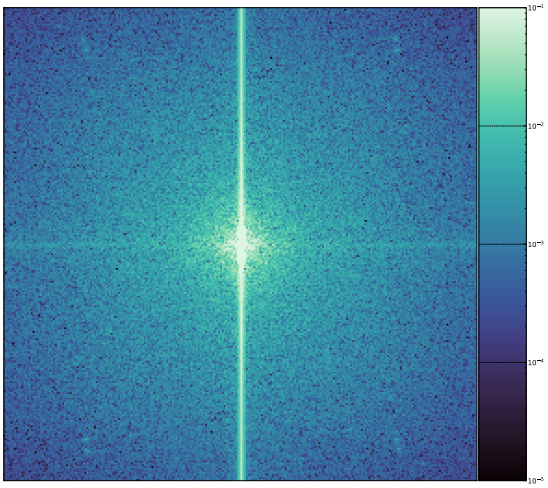
B.1. Resultados de aplicar la Transformada de Fourier Discreta.

B.2. Resultados de aplicar la Transformada de Fourier Discreta con Filtro de Paso Alto.

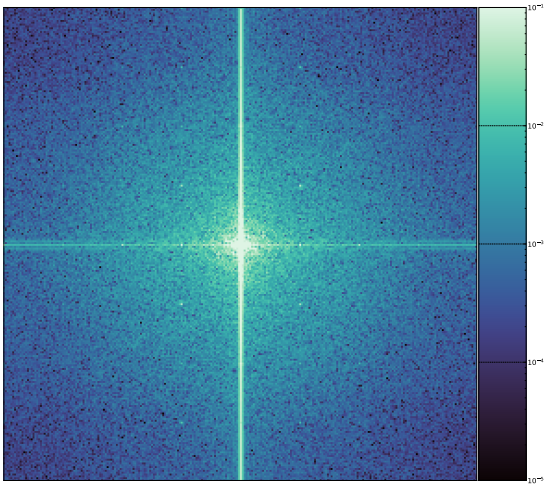
B.3. Resultados de aplicar la Transformada del Coseno Discreta.



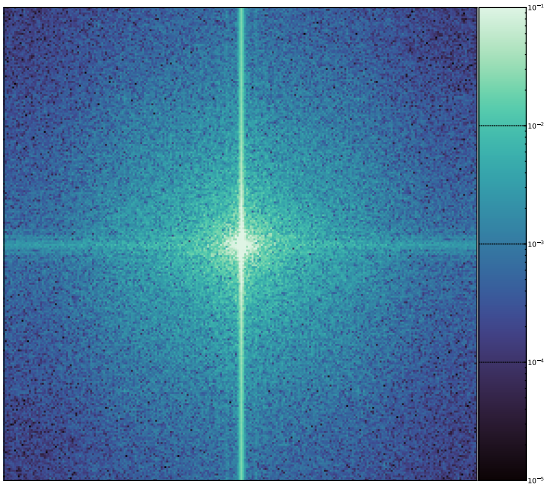
(a) FFT en FFHQ.



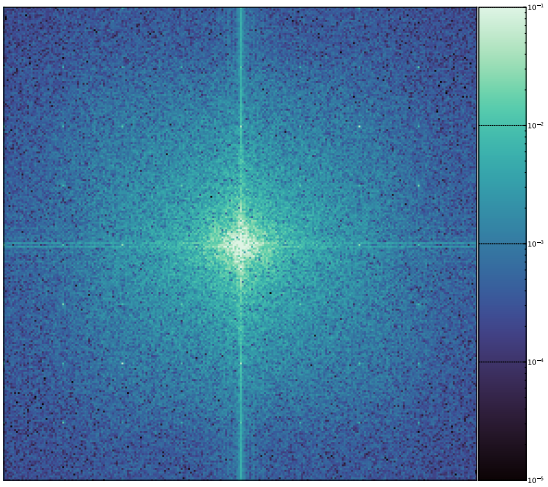
(b) FFT en IMDB-WIKI.



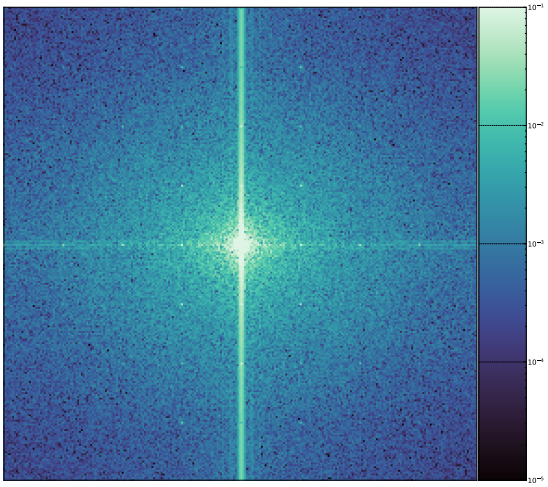
(c) FFT en 'inpainting'.



(d) FFT en 'insight'.

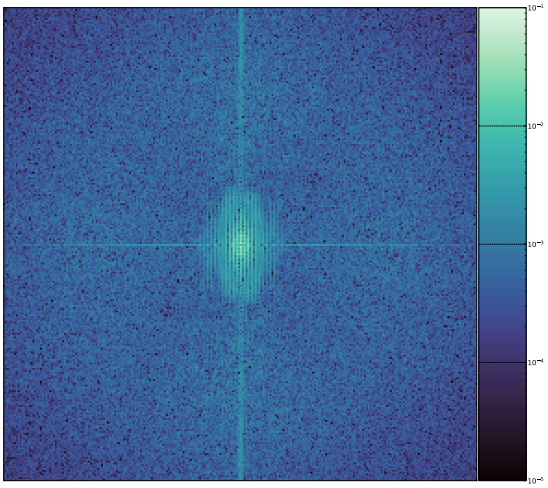


(e) FFT en 'own'.

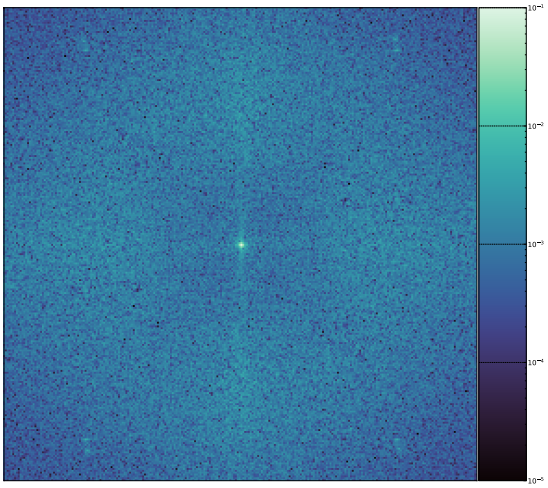


(f) FFT en 'text2img'.

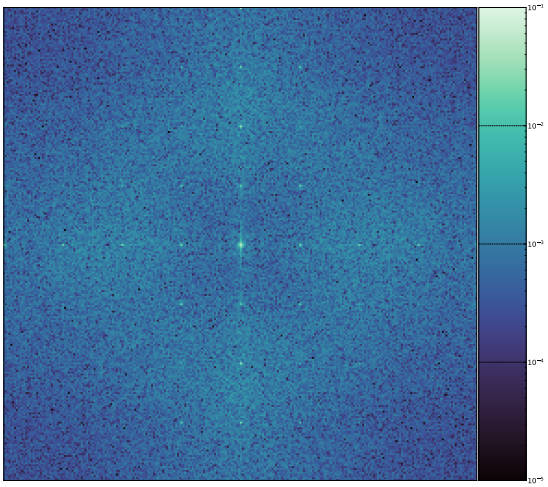
Figura B.1: Resultados para cada conjunto de datos tras aplicar la Transformada Rápida de Fourier (Fast Fourier Transform, FFT).



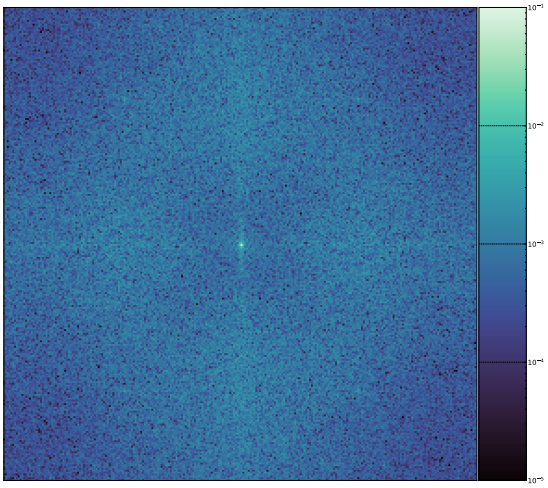
(a) FFT-HP en FFHQ.



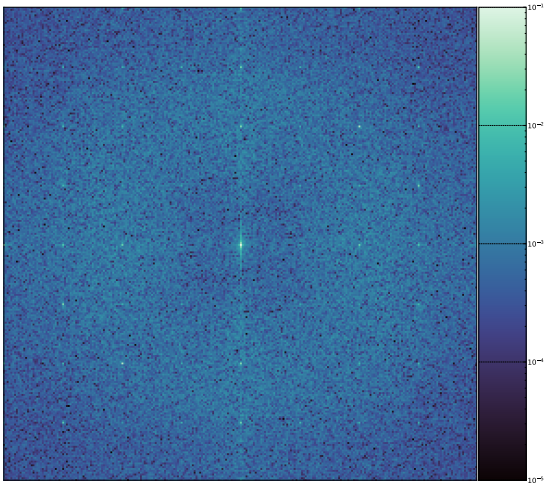
(b) FFT-HP en IMDB-WIKI.



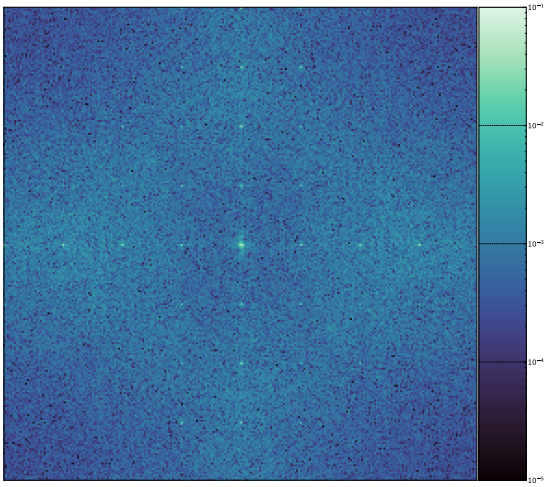
(c) FFT-HP en 'inpainting'.



(d) FFT-HP en 'insight'.

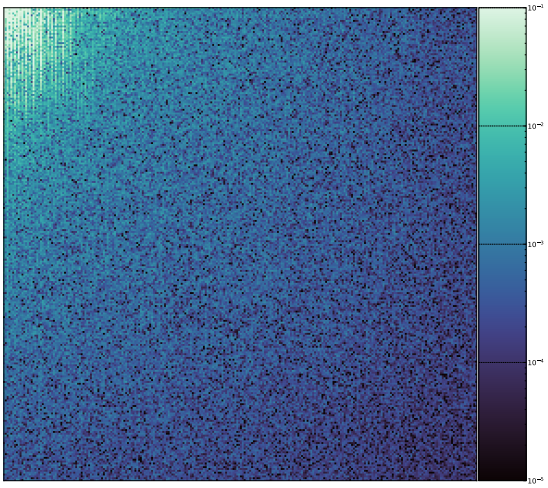


(e) FFT-HP en 'own'.

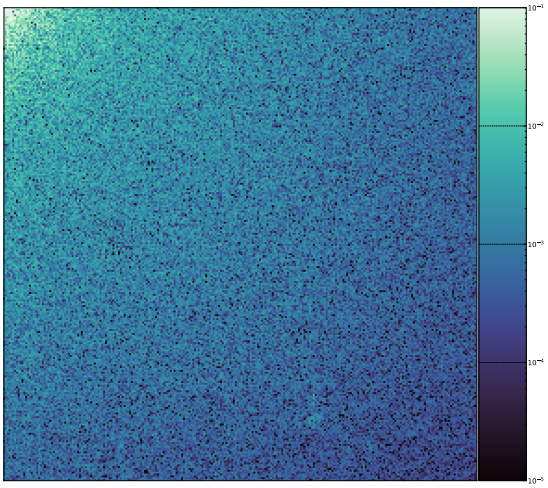


(f) FFT-HP en 'text2img'.

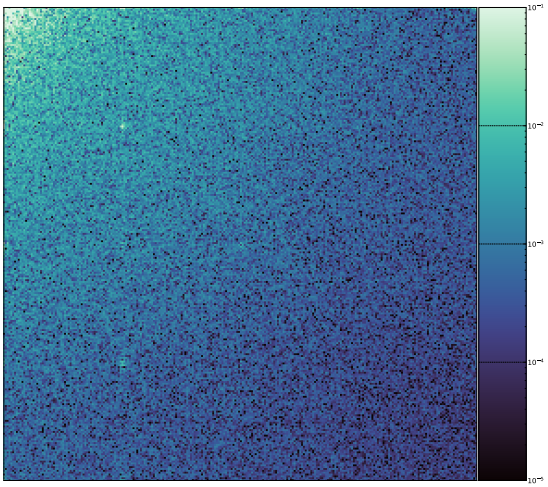
Figura B.2: Resultados para cada conjunto de datos tras aplicar la Transformada Rápida de Fourier con Filtro de Paso Alto (Fast Fourier Transform - High Pass, FFT-HP).



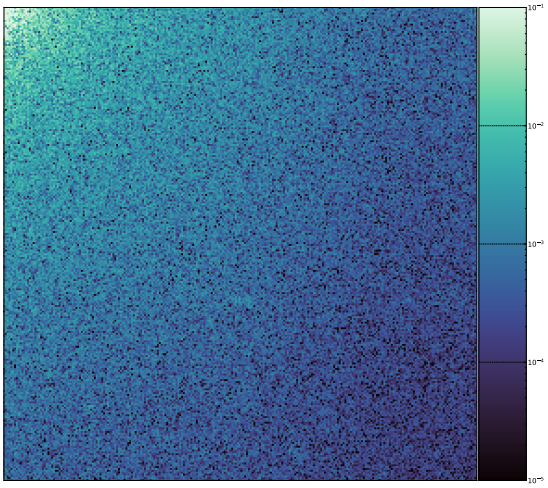
(a) DCT en FFHQ.



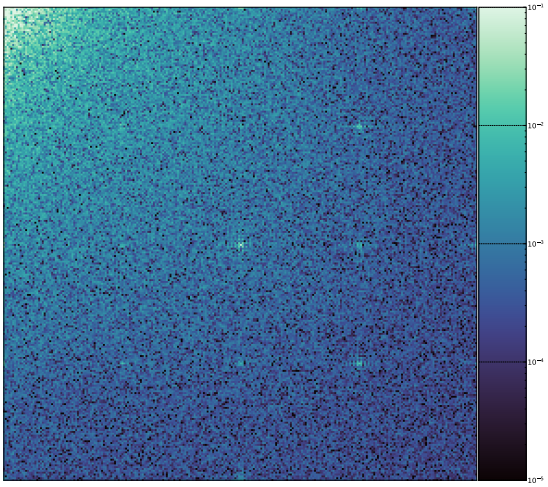
(b) DCT en IMDB-WIKI.



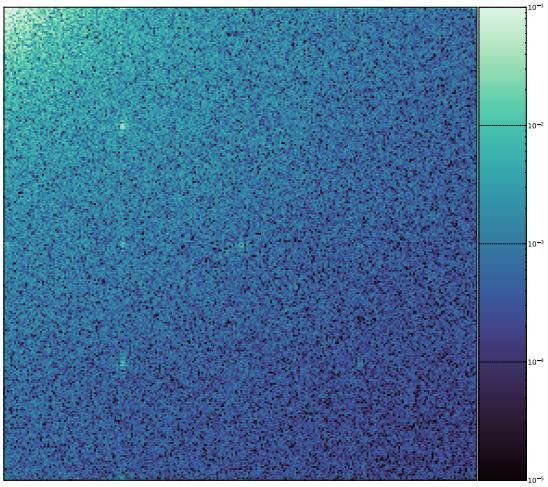
(c) DCT en 'inpainting'.



(d) DCT en 'insight'.



(e) DCT en 'own'.



(f) DCT en 'text2img'.

Figura B.3: Resultados para cada conjunto de datos tras aplicar la Transformada del Coseno Discreta (Discrete Cosine Transform, DCT).