Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física

Maestría en Ciencia de Datos

Modelo Híbrido para la Detección de Deepfakes mediante YOLOv3 y Análisis Espectral con la Transformada de Fourier

---

TESIS para obtener el **GRADO** de
MAESTRO EN CIENCIA DE DATOS

Tesis presentada por:
Uriel Gómez Reyes

Asesor de Tesis:
Dr. Iván Esteban Villalón Turrubiates

Tlaquepaque, Jalisco. Mayo de 2025.

Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física

Formulario de Aprobación de la Maestría en Ciencia de Datos

Título de la Tesis: **Modelo Híbrido para la Detección de Deepfakes mediante YOLOv3 y Análisis Espectral con la Transformada de Fourier**

Autor: **Uriel Gómez Reyes**

Esta tesis ha sido aprobada oficialmente en cumplimiento de todos los requisitos académicos para la obtención del grado de Maestría en Ciencias en Ciencia de Datos.

Asesor de Tesis, **Dr. Iván Esteban Villalón Turrubiates**

Co-Asesor de Tesis, ——

Lector de Tesis, **Mtro. Víctor Hugo Martínez Sánchez**

Lector de Tesis, **Dr. Guillermo Luis Osuna González**

Asesor Academico, **Dra. Rocío Carrasco Navarro**

Tlaquepaque, Jalisco, Mayo de 2025.

Modelo Híbrido para la Detección de Deepfakes mediante YOLOv3 y Análisis Espectral con la Transformada de Fourier

Uriel Gómez Reyes

Resumen

El aumento exponencial en la generación y propagación de deepfakes, sumado al crecimiento de herramientas que promueven su desarrollo, ha originado la necesidad creciente de verificar los contenidos en línea. La posibilidad de que cualquier persona pueda acceder a esta tecnológica sin ninguna dificultad, junto con su asombrosa habilidad para crear rostros hiperrealistas, presenta serios desafíos para su identificación a través de métodos tradicionales.

Este trabajo propone un modelo híbrido que combina el modelo de detección de objetos YOLOv3 sumado con el análisis en el dominio de la frecuencia mediante la Transformada Discreta de Fourier (DFT) y el uso de una Red Neuronal ResNet50 para su clasificación. El modelo utiliza el algoritmo YOLOv3 como una herramienta de visión computacional que permite localizar y extraer las regiones faciales de interés, disminuyendo el ruido del entorno, seguido de un análisis espectral con DFT que permite identificar patrones de alta frecuencia que son característicos de imágenes creadas con herramientas de deepfake. Por último, se entrena un modelo de clasificación con representaciones espectrales de los conjuntos de datos que incluyen imágenes reales y falsas. Esto permite al modelo aprender los patrones de ambos datasets y distinguir entre una imagen real y una falsa.

Los resultados que se obtuvieron como resultado de este trabajo demuestran que el enfoque híbrido propuesto puede lograr una mejora significativa en la precisión de detección respecto a métodos básicos basados exclusivamente en el dominio espacial. Este estudio aporta evidencia sobre la efectividad del análisis de frecuencia como una herramienta más para combatir este tipo de contenido. Combinado con técnicas de visión por computadora, muestra una posible solución para los retos que presenta esta nueva década, dominada por la inteligencia artificial.

Índice general

Índice de figuras

Página

Índice de Tablas

Página

Glosario

Término	Definición
Batch Size	Número de muestras que el modelo procesa antes de actualizar sus pesos durante el entrenamiento. Afecta el rendimiento y la estabilidad del aprendizaje.
Bounding Box	Rectángulo que delimita un objeto detectado dentro de una imagen, comúnmente usado en tareas de detección como YOLO.
CNN o ConvNet (Red Neuronal Convolucional)	Red neuronal especializada en procesar datos con estructura de grilla, como imágenes. Utiliza filtros para detectar patrones visuales y es clave en tareas de visión por computadora y aprendizaje profundo.
Deep Learning (Aprendizaje profundo)	Rama del aprendizaje automático que emplea redes neuronales profundas para reconocer patrones complejos en grandes volúmenes de datos.
Deepfake	Contenido audiovisual generado o alterado con inteligencia artificial para que parezca real, combinando técnicas de *deep learning* y síntesis digital.
DFT (Transformada Discreta de Fourier)	Herramienta matemática que convierte una señal del espacio al dominio de la frecuencia, permitiendo detectar patrones invisibles al ojo humano en el análisis tradicional.
Discriminador	Parte de una red GAN encargada de distinguir entre imágenes reales y falsas generadas por el modelo.
Epoch	Ciclo completo en el que el modelo ha visto todo el conjunto de entrenamiento una vez durante el proceso de aprendizaje.
GAN (Generative Adversarial Network)	Modelo basado en el enfrentamiento entre un generador y un discriminador. Se utiliza para crear imágenes sintéticas realistas.
Generador	Parte de una red GAN encargada de crear imágenes falsas a partir de ruido, con el objetivo de engañar al discriminador.
GPU (Unidad de Procesamiento Gráfico)	Procesador especializado en cálculos paralelos, muy eficiente para entrenar redes neuronales por su arquitectura optimizada para operaciones matriciales.
IA (Inteligencia Artificial)	Rama de la informática que desarrolla sistemas capaces de realizar tareas que requieren inteligencia humana, como ver, hablar o tomar decisiones.
IT (Tecnologías de la Información)	Conjunto de herramientas y procesos relacionados con el almacenamiento, procesamiento y transmisión de datos digitales.
Google Colab	Plataforma en la nube que permite ejecutar código Python en notebooks con acceso gratuito a GPU, ideal para proyectos de inteligencia artificial.
Loss Function (Función de pérdida)	Métrica que mide el error del modelo. Es fundamental para ajustar los pesos durante el entrenamiento mediante retropropagación.
Matplotlib	Biblioteca de visualización en Python utilizada para crear gráficos y visualizar datos durante el desarrollo de modelos.

Término	Definición
NumPy	Librería de Python fundamental para cálculos numéricos con vectores y matrices. Es ampliamente utilizada en ciencia de datos y aprendizaje automático.
OpenCV	Biblioteca de código abierto especializada en procesamiento de imágenes y visión por computadora.
Pipeline	Flujo de trabajo estructurado compuesto por etapas encadenadas que transforman datos de entrada en resultados procesados.
PyTorch	Biblioteca de código abierto en Python para crear y entrenar modelos de aprendizaje profundo. Fue utilizada en este proyecto para implementar la red ResNet-50.
RAM (Random Access Memory)	Memoria volátil que almacena datos temporales mientras se ejecutan programas, esencial para el procesamiento ágil de información.
ResNet-50	Red neuronal convolucional de 50 capas que utiliza conexiones residuales para facilitar el entrenamiento. En este proyecto se usó como clasificador final.
YOLOv3	Algoritmo de detección de objetos en tiempo real que identifica y localiza elementos en imágenes. En esta tesis se empleó para detectar y recortar rostros automáticamente.

1 Introducción

Desde que inicié la maestría en Ciencia de Datos, y a lo largo de mi formación profesional en el área de IT, me ha impresionado la velocidad con la que actualmente la tecnología cambia el mundo en el que vivimos. Lo que antes podría atribuirse a algo de ciencia ficción, el día de hoy forma parte de nuestra vida cotidiana: algoritmos capaces de tomar decisiones clave, ya sea en áreas de salud o en el sector privado, modelos con el poder de redactar textos complejos y entender su significado, sistemas que permiten identificar rostros e incluso redes con la habilidad de generar imágenes tan realistas que es difícil discernir si son reales o no [1].

Bajo este contexto de innovación acelerada y en aumento, surgió un problema que captó profundamente mi atención: las herramientas de deepfakes. La idea de que un algoritmo sea capaz de crear una imagen, un video o una voz y estos puedan ser completamente falsos —pero imposibles de distinguir de nuestra vida real— me provocó tanto asombro como una profunda inquietud. No solo por el nivel técnico alcanzado para desarrollar estas herramientas, sino por sus implicaciones éticas y sociales [2]. Si antes podíamos validar cualquier contenido con la frase "ver para creer", ahora tenemos un problema mucho más grande en nuestras manos: ¿qué podemos hacer cuando ya no podemos confiar en las cosas que vemos?

Esa inquietud personal, alimentada por el creciente interés que producían los medios y mi aprendizaje académico en este fenómeno, fue el punto de partida para poder desarrollar este trabajo de tesis. Con el apoyo de mi asesor, el Dr. Iván, pude enfocar mi trabajo en crear un modelo que pudiera clasificar de forma correcta este contenido multimedia, no solo como un enorme desafío técnico, sino también como una necesidad urgente en un mundo donde lo falso y lo verdadero se entrelazan cada vez más.

En esta investigación busco proponer un enfoque híbrido que combina múltiples herramientas que fui adquiriendo a lo largo de mi estudio en la maestría: El modelo YOLOv3 para su uso en la visión computacional, que me permite localizar y recortar los rostros de las imágenes o videos con alta precisión; el uso de la Transformada Discreta

[1] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020. DOI: 10.1016/j.inffus.2020.07.007

[2] Dario Guera and Edward J. Delp. Deepfake video detection using recurrent neural networks. *AVSS 2018 - IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2020. DOI: 10.1109/AVSS.2018.8639163

de Fourier (DFT), que facilita el análisis espectral en el dominio de la frecuencia, permitiendo detectar patrones invisibles en el dominio espacial, especialmente en imágenes generadas por arquitecturas como StyleGAN [3]; y una red neuronal capaz de clasificar el contenido y aprender de sus características espectrales.

El proyecto se organizó como un modelo híbrido completo, que incluye desde el preprocesamiento y la segmentación facial manual del algoritmo de visión computacional hasta la transformación de esas imágenes al dominio de la frecuencia. Y posteriormente su clasificación utilizando una red neuronal ResNet50. El desarrollo del sistema propuesto se realizó usando únicamente el lenguaje de programación Python, además de la librería de PyTorch, entre algunas otras, y montado en un servicio en la nube. El algoritmo YOLOv3 resultó clave para automatizar la selección de regiones de interés, reduciendo el ruido que pudieran generar los fondos y reduciendo el área de análisis; esto logró mejorar la exactitud del modelo a la hora del entrenamiento. El entrenamiento se llevó a cabo utilizando unos datasets de uso público que incluyen imágenes reales y deepfakes generados por redes como StyleGAN para poder entrenar y probar el modelo.

Los resultados obtenidos son realmente alentadores: demuestran que el análisis en el dominio de la frecuencia puede ofrecer ventajas significativas respecto a métodos tradicionales basados únicamente en el dominio espacial [4], y que el sistema híbrido propuesto representa una alternativa viable para combatir estas tecnologías emergentes. Más allá de las métricas comprobadas, esta tesis también busca abrir el diálogo sobre los riesgos de la manipulación digital, la necesidad urgente de herramientas más robustas, éticas e inteligentes, y la ética detrás de estas herramientas y el uso de inteligencia artificial.

Este trabajo representa la culminación de mi formación en la maestría y tiene como propósito demostrar las habilidades adquiridas a lo largo de mi trayectoria académica y profesional, así como generar un aporte significativo que pueda tener impacto en la sociedad. Porque, queramos o no, esta tecnología seguirá moldeando nuestro futuro; y es nuestra responsabilidad garantizar que su desarrollo y uso se lleven a cabo de forma segura, ética y consciente.

[3] Vito Nicola Convertini, Donato Impedovo, Ugo Lopez, Giuseppe Pirlo, and Gioacchino Sterlicchio. Discrete fourier transform in unmasking deepfake images: A comparative study of stylegan creations. *Information*, 15(711), 2024. DOI: 10.3390/info15110711. URL https://doi.org/10.3390/info15110711

[4] Vito Nicola Convertini, Donato Impedovo, Ugo Lopez, Giuseppe Pirlo, and Gioacchino Sterlicchio. Discrete fourier transform in unmasking deepfake images: A comparative study of stylegan creations. *Information*, 15(711), 2024. DOI: 10.3390/info15110711. URL https://doi.org/10.3390/info15110711

2 Objetivos

Contenido

2.1 Objetivo General

Crear y desarrollar un sistema híbrido para identificar de forma automatizada imágenes deepfake. Este sistema fusiona métodos de visión computacional con análisis espectral, lo que incrementa la exactitud y solidez del procedimiento de detección de deepfakes. Este método se basa en identificar áreas de la cara usando el modelo YOLOv3. Luego, se analizan con la Transformada Discreta de Fourier (DFT) y se clasifican usando una red neuronal profunda llamada ResNet50. Este sistema busca proporcionar una solución automatizada, adaptable y escalable para abordar los problemas que surgen en la detección de contenido sintético producido por redes generativas (GANs).

2.2 Objetivos específicos

1. Investigar y analizar los trabajos y métodos actuales para identificar imágenes modificadas. Esto incluye técnicas tradicionales en el análisis visual de las imágenes y métodos más nuevos que usan representaciones espectrales. Esta revisión me permitirá comprender las restricciones presentes del campo y busca demostrar la justificación de mi trabajo.

2. Elaborar un módulo automatizado de detección facial empleando el modelo YOLOv3, que pueda identificar con exactitud las áreas de interés (rostros) en imágenes o capturas de video. Esta

sección facilitará la disminución del área de análisis, disminuyendo su tiempo de procesamiento, enfocando la tecnología en áreas pertinentes y reduciendo el ruido que pueda generar el fondo. Además, este permitirá el análisis de imágenes que contengan a muchos individuos, abriendo un área de oportunidad que permite una evaluación detallada de cada rostro y la capacidad de examinarlos todos a profundidad, algo que no se puede hacer con otros modelos.

3. Crear una función que convierta imágenes de rostros recortados al dominio de frecuencia usando la Transformada Discreta de Fourier (DFT). El objetivo es identificar patrones de alta frecuencia que se relacionan con imágenes hechas con herramientas de deepfake. Este estudio espectral tiene como propósito identificar posibles patrones que no son perceptibles en otros modelos que están basados en el ámbito espacial; esto permite una posible clasificación usando como herramienta estos patrones detectados.

4. Crear un modelo de red neuronal profunda, específicamente ResNet50, que use representaciones espectrales de las imágenes como entrada y que pueda diferenciar con precisión entre rostros reales y deepfakes. Después de comparar y probar otros modelos, esta red fue seleccionada por su habilidad para gestionar patrones complejos y su efectividad demostrada en labores de clasificación visual.

5. Examinar el desempeño del sistema híbrido sugerido en esta tesis mediante ensayos experimentales, utilizando indicadores como precisión, recall y F1-score, con el objetivo de confirmar su funcionalidad al estar conforme a las condiciones de escenario real. Los resultados de estas evaluaciones ayudarán a verificar si la hipótesis del modelo híbrido puede identificar de manera confiable imágenes que han sido manipuladas, sin necesitar comparaciones con métodos ya existentes. La validación del modelo se basará en sus pruebas sobre conjuntos de datos variados, considerando su capacidad de generalización y su aplicabilidad en condiciones prácticas.

2.3 *Hipótesis*

Se propone el uso de múltiples herramientas que consisten en técnicas de visión computacional a través de YOLOv3, análisis espectral a través de la Transformada Discreta de Fourier (DFT) y clasificación con una red neuronal profunda como ResNet50, que potenciará la exactitud y solidez en la identificación de imágenes deepfake. Con el

objetivo de desarrollar un modelo híbrido, que aspira a optimizar el desempeño de métodos tradicionales enfocados exclusivamente en el análisis espacial y proporcionar un modelo que se pueda optimizar para mejorar su desempeño con el tiempo.

2.4 *Justificación*

El aumento significativo de imágenes creadas de forma artificial, sobre todo con redes generativas antagónicas (GANs), ha dado lugar a un fenómeno que causa preocupación a nivel mundial: los deepfakes. Esta tecnología posee la habilidad para producir rostros falsos con un alto grado de autenticidad o replicar el rostro de un individuo real; estos representan riesgos significativos en diversas áreas, incluyendo la desinformación, la usurpación de identidad, la manipulación en los medios de información y la seguridad digital, entre otras.

Frente a este nuevo panorama, se presenta una urgente necesidad de crear herramientas automatizadas, exactas y escalables que faciliten la identificación confiable de este tipo de contenido. Aunque se han sugerido varios métodos de detección basados en el análisis espacial de las imágenes a través de redes neuronales, muchos de estos tienen limitaciones evidentes al lidiar con manipulaciones más sofisticadas, en las que las diferencias visuales son imperceptibles para el ojo humano y para los clasificadores convencionales.

El sistema híbrido utiliza el algoritmo YOLOv3 para la detección automática de rostros. Este trabajo se fundamenta en su enfoque híbrido que fusiona la visión computacional, el análisis espectral y el aprendizaje profundo, respectivamente. También utiliza el análisis espectral, que convierte las imágenes al dominio de la frecuencia usando la transformada discreta de Fourier (DFT). Se fundamenta en que fusiona la visión computacional, el análisis espectral y el aprendizaje profundo, respectivamente. En su totalidad, esta metodología incrementa la capacidad del sistema para identificar deepfakes con mayor precisión, fiabilidad y adaptabilidad a diversos contextos. Esta solución posibilita abordar el problema desde múltiples ámbitos: la habilidad en el reconocimiento facial, la detección de patrones sutiles en el espectro de frecuencias y una clasificación sólida fundamentada en aprendizaje profundo.

La habilidad de automatizar el proceso facilita la reducción del ruido de fondo y la concentración del análisis exclusivamente en las áreas de interés, potenciando de esta manera la eficiencia y la escalabilidad del sistema. La versatilidad modular de esta estructura permite su eficiencia, y también propone para estudios futuros aplicaciones en tiempo real y sistemas de verificación de contenido en plataformas digitales.

2.5 *Importancia del problema*

Tanto para mí como para muchas personas, la detección de deepfakes ha ganado mucha relevancia en tiempos recientes, no solo por el avance de modelos más potentes para su generación y su accesibilidad, sino también por los riesgos que representan en contextos políticos, sociales y de seguridad. A medida que las herramientas de creación de imágenes deepfake se vuelven más realistas y accesibles a todo el público, la comunidad científica se ha visto con la necesidad de responder con distintos enfoques destinados a detectar contenido creado de forma artificial. A continuación, se describen los principales trabajos desarrollados en esta línea que he recabado en el desarrollo de este trabajo y se contextualiza cómo el presente trabajo ofrece una contribución diferente.

3 Consideraciones éticas, legales y sociales

Contenido

3.1 Justificación ética

Uno de los múltiples objetivos de este trabajo de obtención de grado es impulsar un compromiso ético y un uso responsable de la inteligencia artificial y los algoritmos que se encuentran hoy en día en el mercado, especialmente en el contexto que intenta abordar la tesis, donde crear contenido falso puede tener consecuencias serias tanto para las personas como para la sociedad. Aunque los deepfakes no son solo perjudiciales, también son útiles en muchos sectores, como el cine para efectos del entorno o del actor, la educación como modelo de aprendizaje y el arte digital como herramienta de asistencia. Sin embargo, conllevan riesgos significativos en áreas como la política, los derechos de autor, el robo de identidad y la vida privada de las personas.

Por esta razón, este estudio se sitúa dentro del ideal de un uso responsable, enfocado exclusivamente en la identificación —y no en la creación— de contenido sintético. El objetivo es ofrecer herramientas que ayuden a proteger la verdad de la información, la identidad de las personas y la confianza en el entorno digital. También busca fomentar la necesidad de una regulación e investigación más detallada sobre estas nuevas tecnologías [1].

[1] Carnegie Mellon University. Deepfakes and the ethics of generative ai, 2023. URL https://tepperspectives.cmu.edu/all-articles/deepfakes-and-the-ethics-of-generative-ai

3.2 Uso de datasets públicos y seguros

Todos los datasets utilizados en esta investigación —tanto de imágenes reales como sintéticas con la finalidad de entrenar y probar la red neuronal— han sido seleccionados cuidadosamente para garantizar su legalidad y su disponibilidad pública bajo licencias abiertas, asegurando la parte legal de este trabajo. Ninguna de las imágenes empleadas en el entrenamiento o verificación del sistema se origina de personas que no hayan otorgado su permiso explícito, también con el fin de garantizar su replicabilidad.

Además, los datos se han procesado exclusivamente con el propósito de entrenar modelos de detección, y no han sido reutilizados ni alterados con fines de generación, suplantación o manipulación de identidades.

3.3 Riesgos del uso de deepfakes

A continuación, se describen algunos de los riesgos más relevantes asociados al uso indebido de la tecnología deepfake:

Suplantación de identidad:

La generación de videos falsos mediante técnicas de deepfake puede representar a una persona diciendo o haciendo cosas que nunca ocurrieron. Este tipo de contenido puede utilizarse para dañar la reputación de figuras públicas o individuos comunes, especialmente en redes sociales. Además del riesgo reputacional, existe un impacto financiero y legal potencial, ya que estos videos pueden emplearse con fines de extorsión o fraude [2].

Desinformación:

Los deepfakes también son herramientas poderosas para la propagación de noticias falsas y la manipulación de la opinión pública. Al producir videos o audios que imitan perfectamente a figuras políticas, periodistas o individuos influyentes, se puede inducir al error a vastos sectores de la población [3].

Violencia digital y pornografía no consentida:

Una de las aplicaciones más preocupantes de los deepfakes es la generación de contenido pornográfico no consensuado, donde el rostro de una persona es superpuesto sobre cuerpos ajenos en videos sexuales. Esta práctica afecta en su mayoría a mujeres y celebridades, y constituye una forma de violencia digital, acoso y explotación [4].

Derechos de autor:

El uso de contenido falsificado mediante técnicas de deepfake plantea serias implicaciones en materia de derechos de autor, especialmente cuando se emplea la imagen, voz o estilo artístico de personas reales —como actores, cantantes o artistas visuales— sin su consentimiento [5].

[2] Forbes México. Ia ha incrementando estafas con 'deepfakes', 2023. URL https://forbes.com.mx/ia-ha-incrementando-estafas-con-deepfakes

[3] Diálogo Político. La desinformación de la ia y las elecciones mundiales, 2025. URL https://dialogopolitico.org/edicion-especial-2025-democracia-artificial/la-desinformacion-de-la-ia

[4] BBC Mundo. La crisis del porno deepfake que afecta a las escuelas coreanas, 2024. URL https://www.bbc.com/mundo/articles/c93p53292kyo

[5] Wired España. La huelga de actores de hollywood y la lucha contra la ia, 2023. URL https://shorturl.at/vAKKp

En estos casos, no solo se vulnera el derecho a la identidad y a la autoría, sino que el material suele ser utilizado con fines comerciales, agravando la infracción al monetizar el trabajo o la presencia pública de los afectados.

3.4 *Beneficios de la detección automatizada*

El sistema propuesto puede aportar significativamente en varios contextos:

- Plataformas de redes sociales que deseen moderar contenido manipulado.
- Investigaciones judiciales o forenses.
- Entornos educativos donde se enseñe alfabetización mediática.
- Empresas tecnológicas (como bancos) interesadas en verificar autenticidad audiovisual.

Además, usar técnicas fáciles como la Transformada de Fourier y modelos ya entrenados como ResNet50 ayuda a crear soluciones prácticas que otros investigadores o instituciones pueden reproducir.

3.5 *Transparencia en el desarrollo*

Todo el código desarrollado durante esta tesis ha sido documentado y estructurado para facilitar su comprensión y reutilización. Se ha optado por herramientas de código abierto (PyTorch, OpenCV, YOLOv3) y por una plataforma de desarrollo accesible como Google Colab Pro, con el fin de asegurar la transparencia, colaboración y replicabilidad del sistema propuesto.

3.6 *Aspectos legales*

El avance de los deepfakes plantea nuevos desafíos legales. Aunque las legislaciones aún se están adaptando a esta tecnología, ya existen iniciativas regulatorias:

- **Estados Unidos:** Algunos estados han implementado leyes que exigen identificar los deepfakes en campañas políticas [6].
- **Unión Europea:** La Ley de Inteligencia Artificial considera de alto riesgo a los sistemas generadores de contenido sintético [7].
- **México y Latinoamérica:** Aunque en fases iniciales, se reconoce la necesidad urgente de regular la identidad digital, la privacidad y la desinformación.

[6] Thomson Reuters. Deepfakes: Federal and state regulation, 2023. URL https://www.thomsonreuters.com/en-us/posts/government/deepfakes-federal-state-regulation

[7] European Commission. Proposal for a regulation on a european approach for artificial intelligence. Online, 2021. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

Aunque este trabajo tiene un enfoque técnico, sus resultados pueden ser aprovechados en contextos judiciales como herramienta de verificación audiovisual.

3.7 *Impacto social potencial*

El impacto de los deepfakes en la sociedad va más allá del ámbito técnico: compromete la confianza pública. Su uso malintencionado puede afectar desde procesos electorales hasta relaciones personales.

Esta tesis busca aportar al equilibrio digital ofreciendo soluciones que:

- Protejan a las personas contra la suplantación y el engaño.
- Sean utilizadas por medios y plataformas para combatir la desinformación.
- Fomenten una mayor conciencia sobre los riesgos del contenido manipulado.
- Sirvan como base para sistemas futuros de verificación audiovisual.

Este impacto social positivo constituye una de las motivaciones fundamentales del presente trabajo, que aspira a fortalecer herramientas de defensa digital y confianza pública en un entorno cada vez más vulnerable a la desinformación.

3.8 *Reflexión final*

Las consideraciones éticas, legales y sociales aquí expuestas reafirman que la detección de deepfakes no es solo un reto técnico, sino también un compromiso con la integridad digital. Este trabajo no busca restringir la creatividad ni el desarrollo tecnológico, sino contribuir con herramientas que favorezcan un uso responsable y consciente de la inteligencia artificial. Proteger la verdad, la identidad y la confianza en el entorno digital es una tarea compartida, y esta tesis representa un paso hacia soluciones más seguras, justas y replicables en beneficio de la sociedad.

4 Marco Teórico

Contenido

4.1 Deepfakes y Redes Generativas Antagónicas (GANs)

Los deepfakes son un contenido tanto visual como auditivo que ha sido creado de forma artificial, en la mayoría de los casos utilizando inteligencia artificial. Estos contenidos buscan emular de forma casi imperceptible los rostros, la voz o los gestos de una persona. Este tipo de tecnología ha causado mucho furor en la actualidad. ya sea por su uso en el cine y las redes sociales, o por los riesgos que representan en fenómenos como la desinformación y la suplantación de identidad.

Para generar estos deepfakes se utiliza una de las Redes Generativas Antagónicas (GANs), data de su creador Ian Goodfellow alrededor del 2014 [1]. Estas redes cuentan con dos componentes: un generador, que intenta crear imágenes sintéticas lo más realistas posibles, y un discriminador, que evalúa si las imágenes son reales o falsas. Ambos modelos compiten entre sí en un proceso de aprendizaje que mejora la calidad de las imágenes generadas.

Uno de los modelos más representativos de esta evolución es StyleGAN, desarrollado por NVIDIA, con la capacidad de crear rostros sintéticos de alta fidelidad [2]. A medida que estas redes han evolucionado, también ha incrementado la dificultad de detectar si una imagen es real o falsa.

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014

[2] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019b

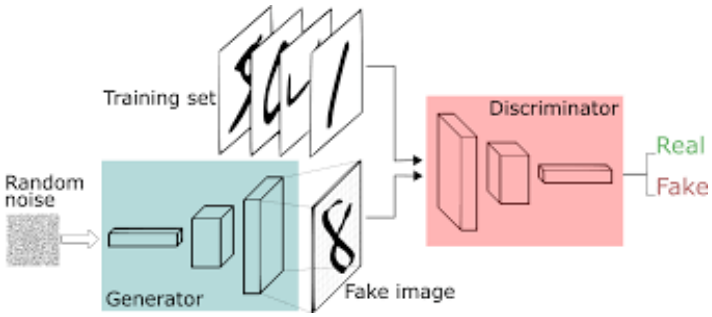

Figura 4.1: Esquema general del funcionamiento de una red GAN. Imagen tomada con fines académicos del sitio https://proyectoidis.org/red-generativa-antagonica-gan/

4.2 *Análisis en el Dominio de la Frecuencia: Transformada Discreta de Fourier (DFT)*

El análisis de una imagen no está limitado por lo que el ojo humano puede ver en el entorno espacial. En la mayoría de los casos, se encuentran patrones importantes en el dominio de la frecuencia, con el uso de herramientas como la Transformada Discreta de Fourier (DFT). Esta técnica transforma una imagen de su forma espacial a una representación que muestra la distribución de sus frecuencias, como lo señala Convertini et al. [3].

Esta herramienta tiene sus raíces en el trabajo de Joseph Fourier, quien demostró que, bajo ciertas condiciones, una función periódica puede representarse como una serie infinita de senos y cosenos. Esta técnica, conocida como Serie de Fourier, es esencial para analizar patrones, identificar detalles ocultos y transformar datos del dominio temporal o espacial al dominio de la frecuencia.

Hablando de los deepfakes, muchos estudios y artículos han demostrado que las imágenes generadas por GANs presentan estructuras sutiles que, aunque imperceptibles al ojo humano, dejan huellas en el espectro de frecuencias, especialmente en las zonas de alta frecuencia. Estos patrones permiten distinguir entre imágenes reales y sintéticas con alta precisión utilizando DFT.

[3] Vito Nicola Convertini, Donato Impedovo, Ugo Lopez, Giuseppe Pirlo, and Gioacchino Sterlicchio. Discrete fourier transform in unmasking deepfake images: A comparative study of stylegan creations. *Information*, 15(711), 2024. DOI: 10.3390/info15110711. URL https://doi.org/10.3390/info15110711

Figura 4.2: Proceso de conversión de una imagen desde su forma original (a color), a escala de grises y finalmente al dominio de la frecuencia mediante la Transformada Discreta de Fourier (DFT).

4.3 Visión por Computadora y Detección Facial: YOLOv3

La finalidad de esta tesis es emplear el algoritmo YOLOv3 para obtener y recortar las áreas faciales de imágenes o fotogramas de vídeos. Este algoritmo posibilita enfocar el análisis espectral en las zonas de mayor relevancia, lo que contribuye a reducir el ruido de fondo y, al mismo tiempo, potencia la eficiencia del proceso de detección.

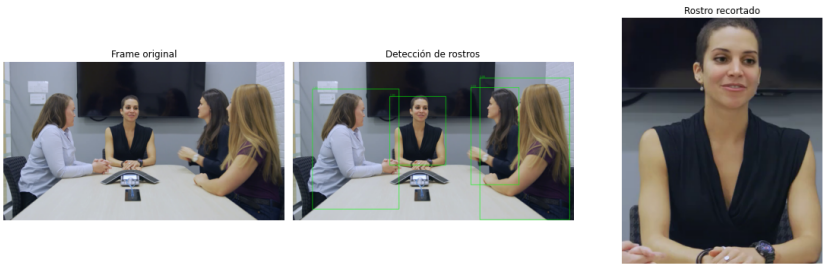

Figura 4.3: Proceso de detección facial utilizando YOLOv3. De izquierda a derecha: fotograma original, detección de rostros mediante bounding boxes, y rostro recortado.

YOLOv3 se introdujo en 2016 por Joseph Redmon como un algoritmo para el uso de identificar objetos en tiempo real [4]. Es importante mencionar que el mismo autor optó por dejar de desarrollar esta tecnología en 2020 por motivos éticos. Mediante una declaración pública, manifestó su inquietud por la utilización militar y de supervisión a gran escala.

[4] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement, 2018. URL https://arxiv.org/abs/1804.02767. arXiv preprint arXiv:1804.02767

4.4 Redes Neuronales Convolucionales (CNN) y ResNet50

En esta tesis, se utilizaron Redes Neuronales Convolucionales (CNNs), estas están diseñadas para manejar datos con una estructura en forma de cuadrícula, tal como sucede con las imágenes o lo que vendría a ser lo mismo, fotogramas en un video. Estas redes se componen de diferentes niveles que llevan a cabo operaciones de convolución, activación y reducción, lo que facilita la obtención de características que van desde elementos básicos como los bordes, hasta patrones más sofisticados como formas, texturas o rasgos faciales, y en nuestro objetivo, el dominio de la frecuencia.

Se utilizó una arquitectura ResNet50, es una red que se distingue por la utilización de bloques residuales [5]. Esta característica le facilita el aprendizaje de funciones de identidad cuando se requieren, lo que es crucial para prevenir el problema del desvanecimiento del gradiente que frecuentemente se presenta en redes muy profundas. Esta habilidad permite que ResNet50 se prepare de manera eficaz, incluso con datos complejos como los que se presentan al distinguir entre rostros auténticos y creados de manera artificial (deepfakes) y proporciona un resultado más exacto.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016a

4.5 Enfoque

Los conceptos presentados en este capítulo establecen la base técnica sobre la cual se construye la propuesta de este trabajo. La combinación de análisis espectral mediante DFT, detección facial con YOLOv3 y clasificación con ResNet50 permite abordar el problema de la detección de deepfakes desde múltiples perspectivas, integrando visión por computadora, teoría de señales y aprendizaje profundo.

4.6 Conjunto de datos utilizado

Para el entrenamiento y evaluación del sistema de detección de deepfakes propuesto, se emplearon tres conjuntos de datos representativos que permiten abarcar distintos escenarios de generación y manipulación de contenido visual. Estos datasets incluyen tanto imágenes reales como sintéticas, así como videos manipulados, lo que garantiza una cobertura amplia de casos reales y falsificados. A continuación, se describen en detalle en la siguiente tabla:

Dataset	**Tipo**	**Fuente**	**Tamaño**	**Uso principal**	**Licencia**
FFHQ	Real	NVIDIA / Flickr	~70K imágenes	Entrenamiento	CC BY-NC-SA
1M Fake Faces	Sintético	A. Reben / StyleGAN	~1M imágenes	Entrenamiento	CC BY-NC 4.0
DFD	Video real/fake	FaceForensics++	~1K videos	Prueba YoloV3	MIT

Tabla 4.1: Comparación entre todos los DataSets utilizados en el sistema propuesto.

5 Estado del Arte

Contenido

5.1 Trabajos previos de interés

Algunos de los primeros trabajos relevantes de los que se recabó información fue el de Li et al. (2018), quienes propusieron detectar deepfakes analizando el parpadeo de personas en videos que fueron alterados [1], centrándose exclusivamente en el área espacial que ofrece únicamente el formato de video. Basaron su hipótesis en el hecho de que, en ese momento, los modelos generativos no podían simular de forma natural el parpadeo humano. Aunque fue un enfoque novedoso, ya que se basaba en la necesidad humana de parpadear, se centraba solo en características visuales visibles y no era suficiente frente a los modelos de deepfake más recientes y avanzados, que pueden imitar este comportamiento con más precisión.

Más adelante, la investigación de Afchar et al. (2018) presentó *MesoNet*, una red neuronal convolucional ligera diseñada para identificar manipulaciones faciales tanto en imágenes como en videos [2]. Su estructura compacta permitía realizar inferencias rápidas con bajo costo computacional, ideal para entornos en tiempo real o con recursos limitados. No obstante, al enfocarse únicamente en el dominio espacial y en los cambios perceptibles, su desempeño se reducía considerablemente al confrontar deepfakes de una calidad muy superior, cuyas modificaciones visuales ya no son tan notorias. Además de esto, presentaba vulnerabilidades frente a imágenes producidas por modelos más sofisticados como StyleGAN, cuyas imperfecciones se ocultan incluso bajo una observación minuciosa.

Por su parte, Rössler et al. (2019) desarrollaron el conocido dataset *FaceForensics++*, ampliamente adoptado como benchmark en la detección de deepfakes [3]. En su estudio se evaluaron múltiples

[1] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. *arXiv preprint arXiv:1806.02877*, 2018. URL https://arxiv.org/abs/1806.02877

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

[3] Andreas Rössler, 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 (ICCV)*, pages 1–11, 2019

arquitecturas, destacando el uso de *XceptionNet*, una red profunda inicialmente creada para clasificación general de imágenes, adaptada aquí al contexto de manipulación facial. Si bien sus resultados demostraron una alta precisión en entornos controlados y con imágenes de buena calidad, el enfoque de su trabajo se limitaba exclusivamente al análisis espacial. Hacía acopio de imágenes recortadas y redimensionadas de los rostros, lo cual reducía su capacidad para detectar manipulaciones más sutiles o parciales.

Trabajos mucho más recientes, como es el caso de Convertini et al. (2024), propusieron utilizar la Transformada Discreta de Fourier (DFT) para localizar huellas espectrales en imágenes generadas por GANs [4]. Este trabajo fue clave para esta investigación, ya que demostró que las imágenes deepfake contienen patrones de frecuencia detectables, parte de lo que se basa nuestra investigación. Ellos hicieron uso de redes neuronales entrenadas sobre este tipo de transformaciones, dando como resultado una mejora a métodos que hacen uso puramente espacial. Sin embargo, su modelo trabajaba únicamente sobre imágenes estáticas, sin considerar las secuencias de video ni integrar algún tipo de proceso automatizado de detección facial para múltiples rostros, lo que limita su aplicabilidad práctica.

[4] Vito Nicola Convertini, Donato Impedovo, Ugo Lopez, Giuseppe Pirlo, and Gioacchino Sterlicchio. Discrete fourier transform in unmasking deepfake images: A comparative study of stylegan creations. *Information*, 15(711), 2024. DOI: 10.3390/info15110711. URL https://doi.org/10.3390/info15110711

En comparación, Nguyen et al. (2019) crearon un pipeline de etiquetado multitarea que fusionaba la segmentación facial, la obtención de características y la clasificación [5]. Aunque incluía una etapa de prelocalización para centrarse en el análisis, su diseño estaba solo enfocado en el aprendizaje espacial y no en el uso de transformaciones para el dominio de la frecuencia. Esto podría hacer que no funcionara en todos los escenarios.

[5] Huy Nguyen, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *arXiv preprint arXiv:1906.06876*, 2019

5.2 *Diferenciación de este trabajo*

El presente trabajo se diferencia de los anteriores en varios aspectos fundamentales. En primer lugar, adopta el análisis en el dominio de la frecuencia como base, al igual que Convertini et al., pero lo extiende mediante una etapa previa de detección facial automática con *YOLOv3* [6]. Esta decisión permite centrar el análisis espectral exclusivamente en las regiones relevantes que se desean —los rostros—, reduciendo el ruido y aumentando la precisión. Igualmente, facilita el análisis de múltiples rostros en una sola imagen o fotograma, lo cual no se contempla en estudios previos de los cuales ellos hacen acopio de imágenes preprocesadas manualmente.

[6] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement, 2018. URL https://arxiv.org/abs/1804.02767. arXiv preprint arXiv:1804.02767

En segundo lugar, mientras otros estudios se apoyan en el uso de redes neuronales básicas, nuestro trabajo utiliza una arquitectura *ResNet50*, reconocida por su eficacia en tareas complejas y su capacidad de generalización [7]. Esto nos permitió elaborar un pipeline sólido que

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016a

utiliza instrumentos desde la detección facial (YOLOv3), a formar parte del preprocesamiento espectral (DFT), hasta la clasificación final con aprendizaje profundo.

Para finalizar, este trabajo se distingue sobre el resto por su diseño modular. Cada parte del sistema propuesto—detección, transformación, clasificación—está claramente separada dentro del modelo, lo que hace más fácil adaptarlo a cualquier sistema, controlarlo en caso de fallos y mejorarlo con diferentes conjuntos de datos o al usar nuevos algoritmos de procesamiento o transformación. La adaptabilidad que propone permite que el sistema en sí no solo sea eficaz, sino también altamente escalable, estableciendo una base firme para futuros estudios derivados del mismo.

6 Marco Metodológico

Contenido

La presente obtención de grado se creó tomando en cuenta un enfoque experimental y cuantitativo, orientado al diseño e implementación del mismo y a la evaluación de un sistema híbrido con la finalidad de poder hacer una detección automatizada de imágenes deepfake. Este sistema integra técnicas de visión por computadora, análisis espectral y aprendizaje profundo, a través de un pipeline modular y escalable. La metodología utiliza herramientas modernas de procesamiento visual y modelos ya entrenados para detectar contenido manipulado con inteligencia artificial.

6.1 Enfoque general del sistema

El sistema híbrido propuesto consta de tres módulos principales:

- Detección facial automática utilizando el modelo YOLOv3.
- Transformación espectral de los rostros detectados mediante la Transformada Discreta de Fourier (DFT).
- Clasificación profunda de las imágenes espectrales mediante una red neuronal ResNet50.

Cada módulo se desarrolló de manera autónoma y luego se incorporó al modelo, lo que facilitó la creación de una arquitectura con una función versátil y de fácil adaptación para replicar el trabajo sugerido o con la intención de generar futuras investigaciones derivadas de este. Este método también posibilita la valoración independiente de cada elemento y su influencia en el desempeño global del sistema, posibilitando su optimización y calibración si se requiere.

6.2 *Herramientas tecnológicas utilizadas*

El sistema fue completamente implementado en el entorno de **Google Colab**, lo que permitió el uso de unidades de procesamiento gráfico (GPU) para acelerar el entrenamiento de modelos, así como la integración con Google Drive para la gestión de archivos y datasets.

Las principales librerías y tecnologías utilizadas en el proyecto fueron:

- Python 3.10 como lenguaje de desarrollo principal.
- Google Colab como plataforma de programación en la nube.
- OpenCV (`cv2`), junto con `numpy`, `os` y `sys`, para implementar YOLOv3, realizar detección facial y manejar imágenes y videos.
- Pillow (PIL), con configuración de tolerancia a imágenes incompletas.
- PyTorch y torchvision para cargar, adaptar y entrenar el modelo ResNet50.
- Scikit-learn para generar reportes de evaluación y métricas.

6.3 *Dataset y preparación de datos*

1. **Flickr-Faces-HQ (FFHQ)** — Rostros reales. Es un conjunto de datos de alta resolución publicado por NVIDIA, inicialmente diseñado para la valoración de modelos generativos adversarios (GAN). Proporciona una extensa variedad visual con fotografías públicas de individuos reales, perfecta para ilustrar la clase .ªuténtica.en actividades de categorización. [1].
2. **1 Million Fake Faces** — Rostros artificiales. Es un conjunto de datos generados por Alexander Reben mediante el uso de la tecnología StyleGAN, que incluye aproximadamente un millón de imágenes completamente artificialessin ninguna conexión con personas reales. Este conjunto de datos en concreto se utilizó específicamente para entrenar al modelo de la red neuronal en su tarea de clasificación para encontrar patrones relacionados con imágenes deepfake. [2].

[1] Tero Karras, Samuli Laine, and Timo Aila. Flickr-faces-hq dataset (ffhq). https://github.com/NVlabs/ffhq-dataset, 2019a. Accedido el 18 de abril de 2025

[2] Alexander Reben. 1 million fake faces dataset. https://archive.org/details/1mFakeFaces, 2019a. Accedido el 18 de abril de 2025

3. **Deep Fake Detection (DFD)** — Evaluación final. Conjunto descargado desde el servidor oficial de FaceForensics, orientado a la evaluación de sistemas de detección de manipulación facial en videos. En esta tesis fue utilizado como etapa final de validación de detección facial automatizada en video [3].

[3] Google AI, Joon Son Chung, and Andrew Zisserman. Deepfake detection challenge dataset (dfd). https://github.com/ondyari/FaceForensics, 2020. Accedido el 18 de abril de 2025

6.4 *Preprocesamiento de datos*

El preprocesamiento de los datos implica una función vital para el sistema propuesto en este trabajo; este representa un conjunto de cambios realizados a las imágenes de origen, en algunos casos, para que sean preparadas correctamente antes de ser examinadas por el modelo. Seguidamente, se presenta un esquema ejemplificativo del proceso de procesamiento:

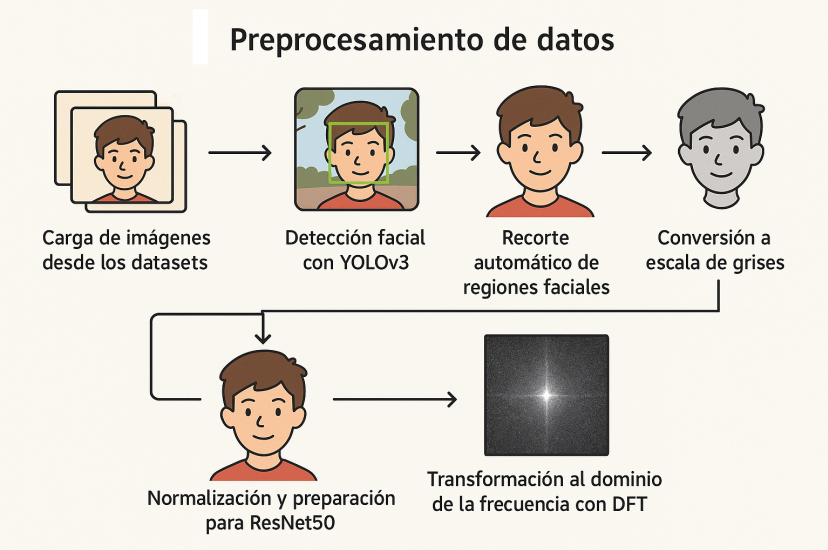

Figura 6.1: Diagrama del flujo de preprocesamiento de datos.

Los pasos incluidos en este proceso son:

- Carga de imágenes desde los datasets.
- Detección facial con YOLOv3.
- Recorte automático de regiones faciales.
- Conversión a escala de grises.
- Transformación al dominio de la frecuencia con DFT.
- Normalización y preparación para ResNet50.

6.5 División del conjunto de datos y entrenamiento

Se utilizó el 80 % de los datos para entrenamiento y el 20 % restante para validación y prueba. El modelo ResNet50 fue ajustado y entrenado sobre las imágenes espectrales resultantes.

6.6 Flujo general del sistema

La Figura 6.2 muestra una representación simplificada del flujo general del sistema propuesto para la detección de deepfakes. En ella se resumen las etapas principales que conforman el proceso, desde la entrada de datos hasta la obtención de resultados clasificados.

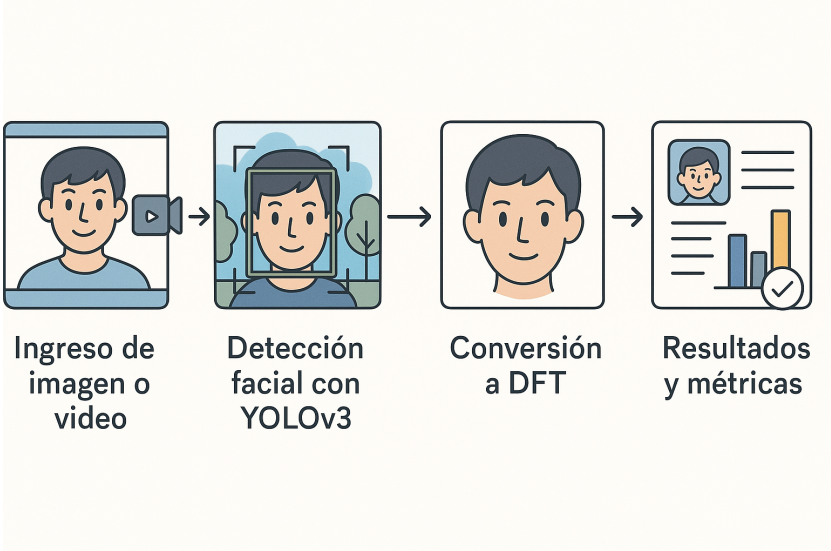

Figura 6.2: Representación simplificada del flujo general del sistema.

El proceso completo incluye las siguientes etapas:

- Ingreso de imagen o video.
- Detección facial con YOLOv3.
- Recorte de rostros.
- Conversión a DFT.
- Clasificación con ResNet50.
- Resultados y métricas.

6.7 Evaluación del modelo

Se aplicaron métricas estándar para evaluar el rendimiento del modelo:

- Precisión.
- Recall.
- F1-score.
- Matriz de confusión.

Estas métricas fueron seleccionadas porque permiten evaluar de forma integral la capacidad del modelo para distinguir entre imágenes reales y manipuladas. Al combinar medidas de acierto, sensibilidad y balance entre errores, se obtiene una visión completa tanto de la eficacia como de la confiabilidad del sistema, especialmente en contextos donde los falsos positivos y falsos negativos pueden tener implicaciones importantes.

6.8 *Justificación de elecciones técnicas*

YOLOv3: Elegido por mi experiencia previa durante la maestría y por su naturaleza de código abierto, lo que permite transparencia total. Aunque existen versiones más nuevas como YOLOv5 o YOLOv8, estas se han vuelto más restrictivas o están controladas por empresas privadas, lo cual va en contra del enfoque ético y abierto que se pretende mantener en esta tesis. Esta perspectiva está alineada con la visión de su creador, Joseph Redmon, quien se retiró del desarrollo del algoritmo por preocupaciones éticas sobre su uso.

ResNet50: Seleccionado por su equilibrio entre profundidad y rendimiento, además de ser compatible con imágenes espectrales de 3 canales. Mi familiaridad con esta arquitectura permitió una mejor optimización y aprovechamiento en la tarea de clasificación de deepfakes.

Transformada Discreta de Fourier (DFT): Fue elegida en conjunto con mi asesor como técnica central desde las primeras etapas del proyecto, al evidenciar su capacidad para exponer patrones de alta frecuencia imperceptibles en el dominio espacial. Su utilidad teórica y práctica la convirtió en el eje de análisis del sistema propuesto.

6.9 *Consideraciones éticas*

Durante el desarrollo de esta tesis se tomaron en cuenta principios éticos fundamentales. Todos los datasets utilizados fueron públicos y distribuidos bajo licencias abiertas, permitiendo su uso académico.

Además, el propósito de este trabajo no es contribuir a la creación o difusión de contenido falso, sino desarrollar herramientas que ayuden a mitigar los efectos negativos de estas tecnologías, contribuyendo a la protección de la integridad digital y la confianza pública. Este proyecto

busca aportar a la verificación de autenticidad de contenido visual en contextos donde la manipulación puede acarrear consecuencias graves.

6.10 Comparación entre conjuntos de datos utilizados

Dataset	Tipo	Fuente	Tamaño	Uso principal	Licencia
FFHQ	Real	NVIDIA/Flickr	~70K imágenes	Entrenamiento	CC BY-NC-SA
1M Fake Faces	Sintético	A. Reben	~1M imágenes	Entrenamiento	CC BY-NC 4.0
DFD	Video real/fake	FaceForensics++	~1K videos	Evaluación final	MIT

Tabla 6.1: Comparación entre conjuntos de datos utilizados en el sistema propuesto.

6.11 Limitaciones del enfoque

Este sistema se centra exclusivamente en rostros dentro de imágenes estáticas y no considera otras modalidades como audio o lipsync en video. Además, conjuntos de datos como FFHQ pueden presentar sesgos demográficos (por ejemplo, prevalencia de rostros de tez clara), lo que podría limitar la generalización del modelo. Estas limitaciones abren oportunidades para futuros trabajos con enfoques más diversos y multimodales.

7 Implementación del sistema

Contenido

7.1 Entorno de desarrollo

7.1.1 Google Colab Pro

La desarrollo práctico de esta tesis se realizó utilizando el entorno **Google Colaboratory Pro (Colab Pro)**; es un software de computación en la nube creado por Google, que permite la implementación de código Python en notebooks de Jupyter directamente desde el navegador.Esto simplifica su acceso y ajuste desde cualquier computadora. Esta herramienta fue creada para la aplicación de proyectos en ciencia de datos, aprendizaje automático y procesamiento de imágenes, evitando la instalación de software localmente, lo que constituyó un beneficio considerable para el avance de este trabajo, frente a la falta de un equipo de trabajo propio.

En este trabajo se tuvo que necesitar la versión de pago **Colab Pro** principalmente porque durante las pruebas iniciales del trabajo con la versión gratuita se alcanzaron rápidamente los límites computacionales, lo que reducía el alcance de mi trabajo. Fue imprescindible para continuar con el proyecto y lograr el alcance previsto sin interrupciones. Colab Pro ofrece ventajas significativas sobre la versión estándar, tales como:

- Acceso prioritario a GPUs de alto rendimiento (como NVIDIA Tesla T4 o P100), esto permitió la realización de operaciones complejas como el análisis espectral.
- Mayor capacidad de memoria RAM, requerida para manejar conjuntos de datos voluminosos y modelos complejos requeridos en el sistema híbrido.
- 100 unidades de cómputo mensuales, lo que permitió extender las sesiones de entrenamiento sin cortes abruptos, asegurando continuidad en el desarrollo del modelo.
- Acceso a la terminal del sistema, permitiendo una personalización avanzada del entorno de ejecución.

Todos los beneficios mencionados anteriormente son de suma importancia, ya que la versión estándar no cuenta con ellos. Estos beneficios impactaron directamente en el proceso de recorte de imágenes y uso del modelo Yolov3, el entrenamiento y uso del modelo

ResNet50 (Clasificador), por su increíble demanda de gestionar el amplio espectro de imágenes que utilizan distintos datasets, como para llevar a cabo el análisis espectral a través de la **Transformada Discreta de Fourier (DFT)** necesaria para lograr su clasificación.

7.1.2 *Lenguaje y librerías utilizadas*

El lenguaje de programación **Python 3.10** fue usado para el desarrollo de este trabajo, famoso por su facilidad, adaptabilidad y extenso respaldo en línea proporcionado por su comunidad. Existen varios factores considerados al seleccionar Python como lenguaje predeterminado, tales como su empleo continuo durante la maestría y su implementación práctica en mi entorno de trabajo. Ya que la familiaridad fue clave para poder llevar a cabo el desarrollo.

Una de las herramientas clave en la implementación del sistema fue **PyTorch**, una biblioteca de código abierto creada por Facebook AI Research (FAIR), ampliamente valorada por su flexibilidad, su facilidad de implementación y su integración eficiente con unidades de procesamiento gráfico (GPUs), que se vio beneficiada al integrar la versión de ColabPro. La experiencia adquirida durante una de las clases de la maestria y su compatibilidad con `torchvision` fueron determinantes en su elección. Poder usar los modelos preentrenados como **ResNet50**, que funge como clasificador en el modelo híbrido, permitió desarrollar modificaciones, ya que esta arquitectura fue adaptada para trabajar con imágenes procesadas en el dominio de la frecuencia.

7.1.3 *Librerías complementarias utilizadas*

Durante el desarrollo del sistema se utilizaron las siguientes librerías y herramientas:

- **Procesamiento y detección de imágenes:** OpenCV (`cv2`), NumPy, os, sys: utilizadas para la manipulación de imágenes, lectura de archivos y detección facial con YOLOv3.
- **Conversión de imágenes:** Pillow (PIL): empleada para la carga segura de imágenes, incluso si se encontraban parcialmente corruptas.
- **Aprendizaje profundo:** PyTorch y torchvision: para entrenar y adaptar el modelo ResNet50. scikit-learn: para calcular métricas de evaluación como precisión, recall, F1-score y matriz de confusión.
- **Visualización y análisis:** Matplotlib y seaborn: para representar resultados, curvas de aprendizaje y comportamiento del modelo.

7.1.4 *Organización del proyecto*

La organización de archivos se estructuró en Google Drive para mantener un entorno modular y fácilmente replicable que pudiera ser controlado y organizado. Las carpetas principales fueron:

- `datasets/`: contiene los conjuntos de datos FFHQ (reales), 1M Fake Faces (sintéticos) y DFD (videos para evaluación).
- `yolo/`: incluye scripts y pesos del modelo YOLOv3.
- `outputs/`: almacena resultados intermedios y finales, como imágenes procesadas, métricas y visualizaciones.
- `models/`: Guarda los modelos entrenados exportados en formato `.pt`, que permitiran su uso posterior a su entrenamiento para pruebas.

Esta arquitectura modular permitió llevar a cabo la evaluación de todo el proceso de forma eficaz y posibilitó su reproducción, siendo uno de sus mayores beneficios su portabilidad y uso en la nube, también fusionando de manera coherente y lógica las fases de detección facial, análisis espectral y clasificación profunda que son los 3 pilares principales propuestos del modelo híbrido.

7.1.5 *Visión general del sistema propuesto*

El diagrama subsiguiente representa el flujo general del modelo híbrido desarrollado para la tarea de obtención de grado. El modelo ilustra todas las fases que atraviesa una imagen, desde su fase de preparación hasta el instante final en el que se clasifica. El sistema comienza con una imagen de entrada, ya sea auténtica o creada con algún método de deepfake, que pasa por el algoritmo de detección facial YOLOv3. Esta etapa permite recortar automáticamente la región facial de interés, reduciendo el ruido del fondo y enfocando el análisis en la zona del rostro que es la de mayor interés.

Una vez recortada el área de interés que es la cara, se aplica la Transformada Discreta de Fourier (DFT), la cual convierte la imagen del dominio espacial al dominio de la frecuencia. Esta representación espectral permite evidenciar patrones de alta frecuencia que son difíciles de detectar a simple vista, pero que pueden ser indicativos de contenido sintético.

El resultado del análisis espectral se emplea como entrada para nuestra red neuronal ResNet-50, que ha sido preentrenada con imágenes reales y deepfakes de nuestros conjuntos de datos. Esta red opera como un clasificador, llevando a una proyección respecto a la autenticidad de la imagen: REAL.º "FAKE", asumiendo que la predicción es precisa.

El modelo híbrido sugerido emplea técnicas de visión por computadora, análisis espectral y aprendizaje profundo para desarrollar un sistema robusto que optimiza el desempeño en tareas de detección, lo que cumple con el propósito general de este estudio y valida la hipótesis propuesta.

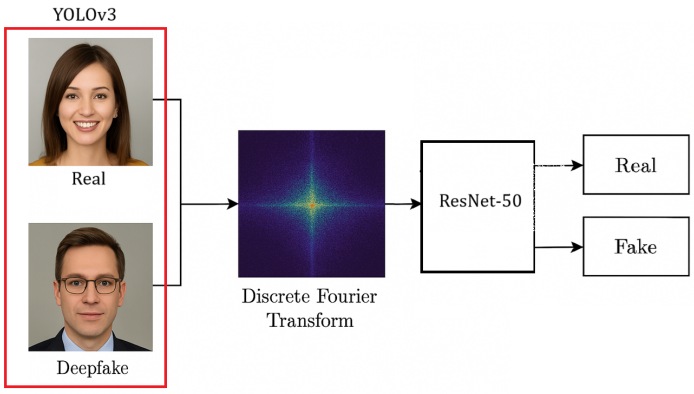

Figura 7.1: Diagrama del sistema Hibrido: Con el modulo de detección facial de YOLOv3, el procesamiento de las imagenes por la transformada de Fourier y modelo de clasificación usando la red Neuronal Resnet50.

7.1.6 *Consideraciones finales del entorno de desarrollo*

El uso de Google Colab Pro como plataforma principal de desarrollo permitió abordar las limitaciones técnicas que implica no contar con un equipo propio de alto rendimiento. El uso de librerías como PyTorch, OpenCV y otras librerías complementarias para el trabajo brindó un entorno flexible, reproducible y versátil para ejecutar cada etapa del sistema híbrido propuesto.

Este entorno no solo permitió ejecutar el sistema completo de forma eficiente, sino que también ofreció una base sólida para su futura expansión o implementación práctica. Es gracias a mi aprendizaje dentro de la institución que fue posible adquirir las técnicas mínimas para poder plantear el proyecto.

7.2 *Implementación de YOLOv3*

YOLOv3 (You Only Look Once, versión 3) fue el modelo seleccionado como el primer eslabón del sistema híbrido propuesto en esta tesis; su principal trabajo es detectar automáticamente los rostros en imágenes y videos [1]. Esta etapa es crucial para el trabajo. Permite reducir el área de análisis al centrarse exclusivamente en las regiones faciales. Esto disminuye el ruido de fondo que podría no ser importante, ayuda a reducir el tiempo de cómputo y mejora la precisión de los módulos posteriores propuestos en el sistema.

[1] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement, 2018. URL https://arxiv.org/abs/1804.02767. arXiv preprint arXiv:1804.02767

7.2.1 *Justificación de la elección*

Se eligió YOLOv3 debido a su capacidad para realizar detección de objetos en tiempo real con una relación óptima entre precisión y velocidad. Aunque existen versiones más recientes como YOLOv5 o YOLOv8, estas se desarrollan bajo estructuras cerradas o con restricciones de licencia a las que, por cuestiones de tiempo, no me aventuré a explorar. En cambio, la versión YOLOv3 es un algoritmo conocido, ampliamente documentado y compatible con bibliotecas como OpenCV, lo que facilitó mucho su elección.

Algo que recalcar es que durante mi formación académica, fue el modelo trabajado y comprendido en mayor profundidad, lo que facilitó su adaptación a este sistema, por lo que decidí no buscar otro tipo de algoritmos de visión computacional. Deseo destacar que esta elección también concuerda con la ética de uso responsable de la inteligencia artificial que trato de abordar también en mi labor, siguiendo el ejemplo de su inventor Joseph Redmon [2], quien optó por dejar el desarrollo debido a inquietudes acerca de su uso inapropiado en situaciones delicadas como la supervisión a gran escala y propósitos militares.

[2] Joseph Redmon. Tweet: I stopped doing computer vision research because... https://x.com/pjreddie/status/1230524770350817280, 2020. Accessed: 2025-04-18

7.2.2 *Fundamentos del algoritmo YOLOv3*

A diferencia de modelos como R-CNN [3], que separan la generación de propuestas de regiones y la clasificación, YOLO realiza ambas tareas en una sola pasada (*one-stage detection*). Divide la imagen de entrada en una cuadrícula y, para cada celda, predice múltiples *bounding boxes* con sus coordenadas, niveles de confianza y clases.

[3] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014

Estas predicciones se ajustan utilizando las siguientes fórmulas, aplicadas internamente por el modelo:

$$b_x = \sigma(t_x) + c_x, \quad b_y = \sigma(t_y) + c_y \tag{7.1}$$

$$b_w = p_w \cdot e^{t_w}, \quad b_h = p_h \cdot e^{t_h} \tag{7.2}$$

Donde:

- t_x, t_y, t_w, t_h son las salidas brutas de la red.
- c_x, c_y son las coordenadas de la celda.
- p_w, p_h son los tamaños de los *anchors*.
- σ es la función sigmoide que normaliza la predicción entre 0 y 1.

Finalmente, se aplica una técnica de supresión no máxima (*Non-Maximum Suppression, NMS*) para eliminar cajas redundantes y conservar la de mayor confianza.

Nota: Aunque estas fórmulas no fueron implementadas manualmente, forman parte del proceso interno de OpenCV al ejecutar `net.forward()`.

7.2.3 *Implementación personalizada*

En este proyecto, se utilizó YOLOv3 con OpenCV, empleando los archivos preentrenados `yolov3.cfg` y `yolov3.weights`, que están relacionados con el dataset **COCO** [4]. Este dataset no tiene una clase específica para `face`, pero sí incluye `person`. Por tanto, se utilizó la clase `person` como guía para estimar la ubicación de los rostros.

El flujo del proceso es el siguiente:

1. Se carga un video y se extraen sus fotogramas.
2. Cada imagen se convierte en un *blob* utilizando `cv2.dnn.blobFromImage`, lo que normaliza la imagen y ajusta su tamaño.
3. Se realiza la inferencia con `net.forward()`.
4. Se filtran las detecciones con confianza mayor a 0.3.
5. Se aplica NMS para eliminar superposiciones.
6. Se extrae el área del rostro a partir del *bounding box* superior.
7. Las regiones faciales se guardan individualmente por carpeta, clasificadas mediante una métrica de superposición llamada *IoU*.

[4] 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 *European conference on computer vision*, pages 740–755. Springer, 2014

7.2.4 *Escalado de imágenes recortadas*

Como parte del preprocesamiento, una vez que se recorta la región facial estimada a partir del *bounding box* generado por YOLOv3, esta imagen resultante se **redimensiona a una escala fija de 224 x 224 píxeles**. Este paso fue necesario por varias razones técnicas:

- Asegurar una entrada uniforme al pipeline de análisis espectral y clasificación.
- Garantizar compatibilidad con las dimensiones de entrada esperadas por la red neuronal ResNet50.
- Evitar distorsiones visuales o espaciales al aplicar la transformada discreta de Fourier (DFT).

Este proceso de escalado posterior al recorte contribuye a estandarizar la representación de todos los rostros procesados, independientemente del tamaño original o del video de origen. Y en parte ayuda al procesamiento, ya que uno de los datasets con los que se entrenó la red neuronal utiliza imágenes de 224 x 224; esta se volvió la resolución de todas las imágenes en el trabajo.

7.2.5 *Seguimiento y organización por individuo*

Encontré un problema al utilizar YOLOv3 que tal vez nadie había planteado antes: cuando un video o imagen contiene más de una persona, si bien se podría analizar toda la imagen, asumí que si la red neuronal estaba entrenada sobre rostros individuales, esto podría generar falsos positivos. Por ello, preferí que YOLO trabajara libremente y que el pipeline procesara cada rostro de forma individual. Esto también permite que, si un solo rostro en una imagen está alterado y el resto no, el sistema tenga la capacidad de detectar el deepfake y señalar específicamente cuál es, en lugar de descartar toda la imagen.

Para esta problemática se implementó un sistema de agrupación por identidad aproximada, utilizando la métrica **Intersection over Union (IoU)**:

$$\text{IoU} = \frac{\text{Área de Intersección}}{\text{Área de Unión}} \tag{7.3}$$

Si el IoU entre una detección nueva y una previamente registrada es mayor a 0.5, se asume que corresponde al mismo rostro y se guarda en la misma carpeta. En caso contrario, se genera una nueva carpeta. Esto es debido a que cuando corría, el algoritmo guardaba los rostros sin un orden en específico y combinaba los rostros en las carpetas; esto me permite almacenar cada rostro en su respectivo archivo y poderlo procesar de forma individual.

Nota: Una limitación de este enfoque es que el tiempo de procesamiento aumenta proporcionalmente al número de rostros detectados. Es decir, a mayor cantidad de rostros en la imagen o video, mayor será el tiempo requerido para procesar cada fotograma, ya que se lleva a cabo un análisis independiente por cada uno de ellos.

7.2.6 *Resultados visuales del sistema*

En la siguiente figura se observa un fotograma procesado con YOLOv3, donde se muestran las cajas de detección y la clase identificada. Como podemos ver en la imagen, al encontrarse dos personas, genera un recuadro encima de las dos figuras.

Detección de rostros en video:

Figura 7.2: Frame original y resultado de detección facial con YOLOv3 sobre un video con dos personas.

Carpeta generada con los rostros extraídos:

Se implementó un sistema donde se guardan los rostros recortados de forma automática; en este caso se ven los fotogramas individuales de una sola de las personas durante el vídeo.

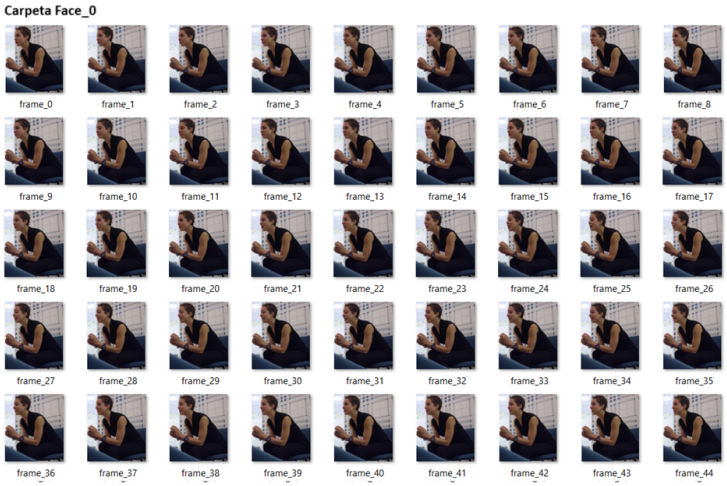

Figura 7.3: Visualización de los fotogramas extraídos y almacenados en una carpeta específica; se puede observar cada uno de los fotogramas que corresponde a esa persona.

7.2.7 *Esquema de arquitectura y pipeline de detección facial*

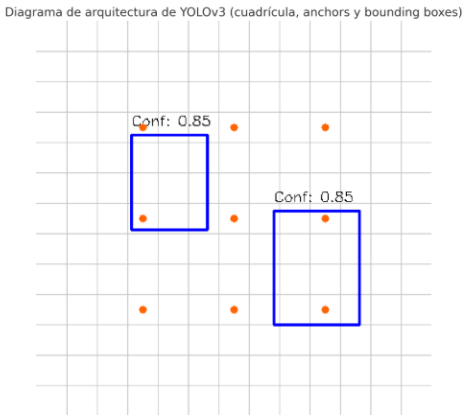

Figura 7.4: Esquema simplificado del funcionamiento interno de YOLOv3: detección de bounding boxes sobre una cuadrícula. Muestra de forma resumida el funcionamiento del sistema de detección facial.

7.2.8 *Trabajo futuro con YOLOv3*

Si bien YOLOv3 ha demostrado ser un algoritmo muy eficaz para la detección de rostros en imágenes y fotogramas en los videos dentro del sistema propuesto, su rendimiento puede ser superado por versiones más recientes. Una línea clara de trabajo futuro para esta parte del módulo podría consistir en la exploración e integración de arquitecturas modernas de este mismo algoritmo, buscando que cumplan con la parte

ética. Estas han demostrado mejoras notables tanto en precisión como en velocidad de inferencia, especialmente en contextos con limitaciones de hardware o en escenarios de procesamiento en tiempo real.

También se podría estudiar cómo usar métodos de fine-tuning en modelos que ya están entrenados para tareas específicas de identificación facial en situaciones más complicadas, como poca luz, diferentes características demográficas o múltiples rostros en la misma escena que desafían al modelo.

Otra mejora importante sería la implementación de mecanismos de seguimiento facial (face tracking) a través de secuencias de video, permitiendo una detección más robusta y coherente a lo largo del tiempo y posibilitando la posibilidad de un modelo OnDemand. Esto reduciría falsos positivos y podría mejorar la precisión general del sistema cuando se aplique a contenido audiovisual extendido en vez del entrenamiento con videos cortos.

7.2.9 *Reflexión final sobre YOLOv3*

La incorporación de YOLOv3 como módulo inicial del sistema híbrido fue clave para acotar el análisis a las regiones faciales de interés, reduciendo el ruido de fondo y mejorando el rendimiento general del pipeline. A pesar de no ser la versión más reciente, su estabilidad, documentación y compatibilidad con bibliotecas como OpenCV permitieron una implementación eficaz y reproducible. Este módulo no solo facilitó una segmentación precisa de los rostros, sino que también sentó las bases para que el sistema trabajara de forma escalable y ordenada, procesando múltiples rostros con autonomía.

En resumen, YOLOv3 resultó ser una herramienta adecuada para los fines de esta tesis, y su integración representa un ejemplo claro de cómo un modelo bien comprendido y éticamente utilizado puede seguir teniendo un impacto relevante en soluciones actuales de visión por computadora.

7.3 *Transformada discreta de Fourier (DFT)*

Este análisis es especialmente relevante dado que, a medida que los modelos de deepfake se vuelven más sofisticados y robustos, las diferencias visuales entre imágenes reales y sintéticas serán cada vez más difíciles de distinguir. En este contexto, la DFT se utiliza para revelar patrones característicos en el espectro de frecuencia de las imágenes deepfake con las cuales contamos que podrán ser de ayuda a la hora de hacer la clasificación.

En el presente capítulo se describe detalladamente el proceso de aplicación de la Transformada Discreta de Fourier dentro del sistema

propuesto, así como la justificación de su uso en el marco del análisis espectral para detección de contenido manipulado.

Figura 7.5: Conversión de una imagen FFHQ a escala de grises y su transformación al dominio de la frecuencia.

7.3.1 *Introducción al uso de DFT en deepfakes*

Este análisis es particularmente significativo ya que, conforme los modelos de generación se hacen más avanzados, las diferencias físicas entre imágenes auténticas y artificiales serán cada vez más arduas de diferenciar. En este escenario, se emplea la DFT para descubrir patrones distintivos en el espectro de frecuencia de las imágenes deepfake [5].

[5] Vito Nicola Convertini, Donato Impedovo, Ugo Lopez, Giuseppe Pirlo, and Gioacchino Sterlicchio. Discrete fourier transform in unmasking deepfake images: A comparative study of stylegan creations. *Information*, 15(711), 2024. DOI: 10.3390/info15110711. URL https://doi.org/10.3390/info15110711

7.3.2 *Fundamento matemático básico*

La transformada discreta de Fourier toma una imagen (una función de dos dimensiones de la intensidad de los píxeles) y la convierte en una representación que muestra la intensidad de varias frecuencias que están presentes en la imagen. La fórmula general es:

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \cdot e^{-j2\pi(ux/M+vy/N)} \tag{7.4}$$

Donde:

- $f(x,y)$: intensidad del píxel en la posición (x,y)
- $F(u,v)$: componente de frecuencia correspondiente
- M, N: dimensiones de la imagen

Esta operación fue ejecutada mediante la función `np.fft.fft2()` de la biblioteca NumPy. El resultado se centra con `np.fft.fftshift()` y puede visualizarse en una imagen con intensidades brillantes que representan frecuencias dominantes [6].

[6] NumPy Developers. Numpy fft module documentation. https://numpy.org/doc/stable/reference/routines.fft.html, 2024. Último acceso: abril de 2025

Nota: Todo el cálculo fue automatizado utilizando funciones de NumPy, lo que permitió integrar esta transformación fácilmente en el pipeline general del sistema.

7·3·3 *Implementación dentro del sistema*

Las imágenes faciales extraídas automáticamente por YOLOv3 fueron convertidas a escala de grises para posteriormente ser transformadas al dominio espectral utilizando DFT. El flujo de procesamiento fue:

1. Conversión a escala de grises.
2. Aplicación de DFT con `np.fft.fft2()`.
3. Centrado del espectro con `np.fft.fftshift()`.
4. Normalización.
5. Reconversión a formato RGB duplicando el canal espectral para ResNet50.
6. (Adicionalmente) Extracción de espectro 1D (perfil de potencia radial) para análisis gráfico.

7·3·4 *Justificación de reconvertir el espectro a RGB*

Aunque la Transformada Discreta de Fourier (DFT) genera una imagen en escala de grises que representa la magnitud del espectro de frecuencias, que es lo que necesitamos para nuestro modelo, fue necesario reconvertir este resultado al formato RGB para hacerlo compatible con la arquitectura de la red neuronal ResNet50. Este modelo tuvo la necesidad de ser preentrenado con imágenes de tres canales (RGB), por lo que espera tensores de entrada con dicha estructura.

Para lograr esta compatibilidad sin distorsionar la información espectral, se replicó el canal de intensidad (grises) en los tres canales de color, generando así una imagen RGB sintética. Afortunadamente, el proceso que estamos realizando no provoca ningún tipo de pérdida de información, dado que no afecta los valores iniciales, sino que simplemente duplica la matriz de intensidad en cada uno de los canales correspondientes para poder funcionar.

En cuanto al procesamiento, esta conversión implica una carga de trabajo mínima, dado que consiste en una operación sencilla de duplicación de los canales de datos. No obstante, posibilita el uso directo de modelos ya entrenados, evitando el gasto considerablemente superior de rediseñar o entrenar una red neuronal desde el inicio para poder manejar imágenes en escala de grises.

7.3.5 *Visualización Imágenes reales vs. deepfakes*

La siguiente figura muestra el resultado del análisis espectral sobre dos imágenes, una real y una sintética generada por GANs. En ambas filas se presentan tres elementos:

- La imagen original (real o falsa).
- Su representación en el dominio de la frecuencia usando la Transformada Discreta de Fourier (DFT).
- El espectro de potencia unidimensional (1D Power Spectrum).

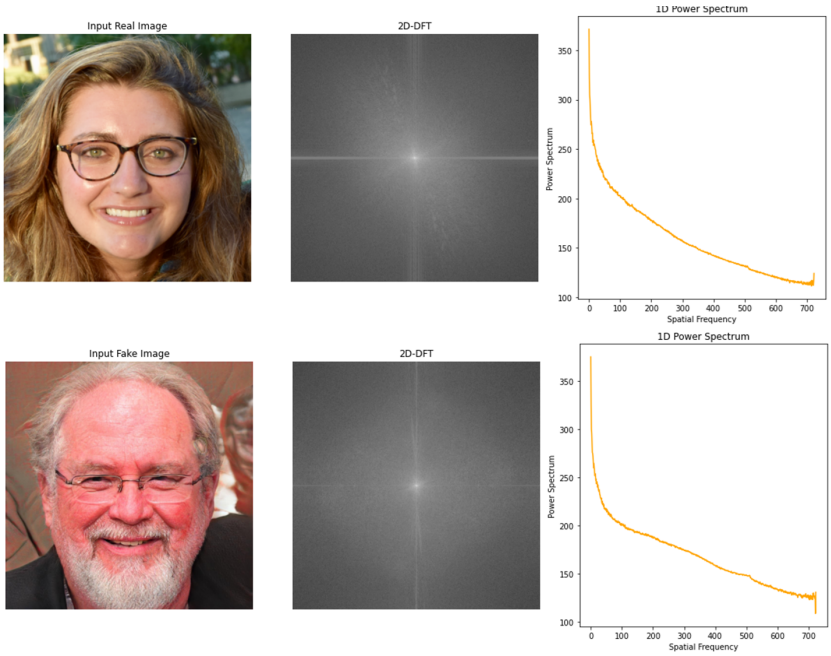

Figura 7.6: Comparación entre imagen real (fila superior) y deepfake (fila inferior) con su respectiva representación espectral y curva de potencia 1D.

En la imagen real sacada del dataset de imágenes reales, el espectro presenta una caída más suave en las frecuencias altas, mientras que en la imagen sintética sacada de su respectivo dataset *1 Million Fake Faces* [7] se observan patrones de frecuencia más marcados y ciertos picos irregulares. Esto sugiere que los deepfakes, aunque visualmente creíbles, dejan huellas que pueden ser reveladas en el análisis frecuencial, especialmente en el comportamiento de las frecuencias altas.

[7] Alexander Reben. 1 million fake faces. https://archive.org/details/1mFakeFaces, 2019b. Dataset generado con StyleGAN, licencia CC BY-NC 4.0

7.3.6 *Limitaciones del análisis espectral*

Aunque el análisis en el dominio de la frecuencia aporta ventajas importantes al detectar patrones estadísticos ocultos en imágenes deepfake, también presenta limitaciones que deben considerarse al momento de evaluar su aplicabilidad práctica:

- **Falta de localización espacial:** La DFT proporciona información global sobre la imagen, pero no permite identificar con precisión en qué región del rostro ocurren las anomalías, lo que limita el análisis interpretativo o explicativo.
- **Sensibilidad al ruido y la compresión:** La calidad de la representación espectral puede degradarse significativamente ante la presencia de ruido, compresión JPEG, o distorsiones introducidas durante la transmisión o manipulación de la imagen [8].
- **Requiere escalado uniforme:** Para que las frecuencias representadas sean comparables entre muestras, es necesario escalar las imágenes a un tamaño estándar. De lo contrario, la variabilidad espacial afecta directamente la validez del análisis espectral.

[8] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your step: Learning graphical representations for deepfake detection using frequency analysis. *arXiv preprint arXiv:2005.02791*, 2020

Estas limitaciones no anulan el valor del enfoque, pero deben considerarse si se busca extender el sistema a entornos más adversos o aplicarlo en tiempo real.

7.3.7 *Posibles mejoras futuras*

Una extensión lógica de este trabajo sería utilizar técnicas mixtas como:

- **STFT (Short-Time Fourier Transform)**: Permite analizar las frecuencias en ventanas móviles dentro de la imagen.
- **Transformada Wavelet**: Brinda tanto localización espacial como frecuencial con mayor precisión y adaptabilidad.

Estas técnicas podrían complementar el análisis de DFT y ofrecer una visión más detallada en imágenes de mayor resolución, y sería importante ver su funcionamiento en los trabajos futuros que se generen después.

7.3.8 *Reflexión final sobre el uso de DFT*

La integración de la Transformada Discreta de Fourier dentro del sistema híbrido no solo aportó un enfoque complementario que puede ser usado en el análisis visual tradicional, sino que permitió descubrir patrones sutiles presentes en el dominio de la frecuencia que resultan invisibles a simple vista. Este enfoque evidenció que, incluso cuando los deepfakes logran engañar al ojo humano, aún dejan huellas estructurales detectables por medio de técnicas espectrales, reafirmando así el valor de la DFT como una herramienta poderosa en la lucha contra la manipulación digital.

7.4 Clasificación con ResNet50

Después de transformar las imágenes al dominio de la frecuencia usando el método anterior de la Transformada Discreta de Fourier (DFT), el siguiente paso del sistema híbrido propuesto es clasificar las imágenes como reales o deepfakes. Para lograr este objetivo, se necesitó el uso de una red neuronal convolucional basada en la arquitectura ResNet50.

7.4.1 ¿Qué es ResNet50?

Su nombre viene del inglés *Residual Networks*, en corto ResNet, y el número adjunto en el nombre se debe a que hace uso de 50 capas de profundidad. Esta es una red convolucional creada en 2015 por el equipo de Multimedia de Microsoft Bing [9]. Se distingue de otras redes al introducir el concepto de aprendizaje residual, contribuyendo a solucionar uno de los retos más importantes en el campo del deep learning: el reto de entrenar de forma eficaz a medida que la profundidad de la red aumenta.

[9] Microsoft Research. Microsoft vision model resnet-50 combines web-scale data and multi-task learning to achieve state-of-the-art. https://shorturl.at/8MYeZ, 2023. Accessed: 2025-04-18

A diferencia de redes anteriores, que intentaban aprender directamente una función de mapeo $\mathcal{H}(x)$, ResNet sugiere que es más fácil para la red aprender la diferencia o el *residuo* entre la entrada y la salida esperada, es decir, $\mathcal{F}(x) = \mathcal{H}(x) - x$. Esto lleva al siguiente bloque residual:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}) + \mathbf{x} \tag{7.5}$$

donde $\mathbf{x}$ es la entrada, $\mathcal{F}(\mathbf{x})$ representa una serie de operaciones convolucionales, y $\mathbf{y}$ es la salida del bloque residual.

Arquitectura general de ResNet50

Esta red está compuesta por bloques de convoluciones de tipo 1x1, 3x3 y 1x1, organizados en módulos residuales que se repiten múltiples veces. La arquitectura sigue una estructura jerárquica en la que se reduce progresivamente la dimensión espacial mientras se incrementa la profundidad de cada uno de los canales.

A continuación se muestra una representación esquemática del flujo de datos en ResNet50, desde la entrada hasta la salida final:

Justificación y ventajas

La necesidad de un modelo preentrenado como ResNet se originó al observar que añadir más capas a una red no siempre optimiza su desempeño, e incluso puede deteriorarlo. ResNet resolvió esto permitiendo que las capas adicionales no perjudicaran el modelo si no aportaban aprendizaje, gracias a sus conexiones residuales. Adjunto un

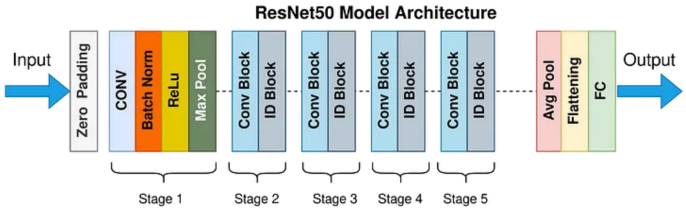

Figura 7.7: Esquema de la arquitectura ResNet50 con bloques residuales, incluyendo las dimensiones de cada etapa.https://shorturl.at/PH6w5. Esta imagen no es de autoría propia.

documento más detallado de su investigación y su desarrollo [10].

Agregue la Tabla 7.4.1; aquí se ve como el modelo ResNet-50 desarrollado por Microsoft supera a otras alternativas populares en varias tareas de clasificación. Siento que este tipo de comparativas es importante para demostrar mi elección con este modelo.

[10] 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 (CVPR)*, pages 770–778. IEEE, 2016b

Tabla 7.4.1.
Evaluación del modelo Microsoft ResNet-50 frente a modelos populares en tareas de visión artificial.
Fuente: Microsoft Research .

Microsoft Research. Microsoft vision model resnet-50 combines web-scale data and multi-task learning to achieve state-of-the-art. https://shorturl.at/8MYeZ, 2023. Accessed: 2025-04-18

Conjunto de datos	Microsoft ResNet-50	Google Big Transfer	CLIP (OpenAI)	PyTorch ResNet-50
CIFAR-10	92.64	92.51	87.85	82.23
CIFAR-100	76.05	79.84	67.02	61.36
STL-10	98.10	98.71	97.20	96.32
SVHN	72.64	64.22	64.33	52.05
Cachorro	82.20	82.75	68.38	38.79
Flores-102	99.28	99.38	95.23	77.62
ImageNet	73.85	72.83	57.00	75.63
Promedio	**84.97**	**84.32**	**76.72**	**69.14**

7.4.2 *Implementación de ResNet50*

Para esta tesis se utilizó una versión preentrenada de ResNet50 proporcionada por la librería `torchvision.models` de PyTorch [11]. Este modelo fue entrenado previamente con el dataset ImageNet, permitiendo aplicar aprendizaje por transferencia a nuestra tarea de clasificación binaria.

[11] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 8024–8035, 2019

Preparación del dataset

- **Escalado:** Las imágenes espectrales fueron redimensionadas a 224x224 px, como requiere la arquitectura.
- **Normalización:** Se utilizó la normalización estándar de ImageNet:

```
transforms.Normalize(mean=[0.485, 0.456, 0.406],

std=[0.229, 0.224, 0.225])
```

- **División de datos:** 80

Adaptación del modelo

La capa final fue reemplazada por una capa densa con una salida y activación sigmoide:

```
model.fc = nn.Sequential(
    nn.Linear(2048, 1),
    nn.Sigmoid()
    )
```

Entrenamiento

- **Épocas:** 20
- **Batch size:** 32
- **Optimizador:** Adam con learning rate 0.0001
- **Función de pérdida:** Binary Cross Entropy Loss

7.4.3 *Versionamiento y ajustes*

Durante la elaboración de este proyecto de tesis, el código experimentó varias versiones, como es habitual en un trabajo de esta naturaleza. En un principio, se empleó un modelo más simple que tenía menos capas, pero se observó que no podía reconocer bien patrones sutiles en los espectros de Fourier, que eran el objetivo principal del trabajo. Luego, se inició ResNet18 para evaluar su capacidad, y después se elevó a ResNet50. Al observar una notable mejora en la capacidad de generalización, se optó por subir a ResNet50. Esta fue la red que se utilizó en todo el desarrollo.

Además, se probaron distintos tamaños de imagen, desde 128x128 hasta 224x224, y se concluyó que este último ofrecía el mejor equilibrio entre costo computacional y precisión, aun a sabiendas de que este factor podría ayudar o no al desarrollo del trabajo.

También se experimentó con el uso de la función de activación Softmax en lugar de Sigmoid, pero dado que la tarea es binaria y el modelo produce una sola salida, Sigmoid fue más apropiada.

7.4.4 *Uso de PyTorch en el proyecto*

PyTorch fue elegido por su flexibilidad, documentación amplia y compatibilidad con GPUs [12]. Se aprovechó `torchvision.models` para cargar ResNet50 preentrenada:

[12] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 8024–8035, 2019

```
resnet = models.resnet50(pretrained=True)
```

Y se adaptó para clasificación binaria:

```
resnet.fc = nn.Sequential(
    nn.Linear(2048, 1),
    nn.Sigmoid()
)
```

El entrenamiento se realizó con GPU Tesla T4 de Google Colab Pro.

7.4.5 *Conclusiones y futuras mejoras*

La integración de ResNet50 como clasificador profundo fue un componente esencial del sistema híbrido. La capacidad inata de la red neuronal ResNet50 de obtener características complejas del dominio espectral permitió la detección de los patrones requeridos que, en el contexto de otros modelos, no serían identificables por clasificadores superficiales. ResNet50, junto con otras herramientas del proceso de preprocesamiento del sistema YOLOv3 y DFT, alcanzó el rendimiento esperado y evaluaciones prácticas sobre la efectividad del método propuesto en esta investigación.

No obstante, uno de los principales desafíos en esta etapa del trabajo fue el largo tiempo que requería entrenar la red, incluso al usar un GPU mejorado. La complejidad del modelo y el volumen de datos procesados aumentaron la demanda de consumo. En consecuencia, para el desarrollo de futuras investigaciones o para el seguimiento de este mismo trabajo, se propone ampliamente explorar alternativas como lo pueden ser EfficientNet, MobileNetV3 o incluso redes transformer como ViT (Vision Transformer), que podrían ofrecer una mejor relación entre precisión y velocidad de entrenamiento. Estas posibles mejoras permitirían escalar el sistema a escenarios en tiempo real o con recursos computacionales más limitados.

De igual manera, se recomienda aumentar la cantidad y diversidad de los datasets de entrenamiento para mejorar la generalización del modelo frente a variaciones reales del entorno, condiciones de luz y diversidad demográfica [13].

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019b; and Alexander Reben. 1 million fake faces. https://archive.org/details/1mFakeFaces, 2019b. Dataset generado con StyleGAN, licencia CC BY-NC 4.0

8 Resultados obtenidos

8.1 Rendimiento del sistema

Durante la etapa de entrenamiento del modelo ResNet50, se utilizaron 40,000 imágenes, de las cuales 8,000 se reservaron para validación, cumpliendo el 80/20 de la división de los datos. La mitad del conjunto de datos pertenece al dataset de imágenes falsas y la otra mitad al conjunto FFHQ-dataset de imágenes reales. El flujo completo, incluyendo preprocesamiento, entrenamiento y evaluación, se realizó utilizando la herramienta de Google Colab Pro y tomó aproximadamente 6 horas continuas de cómputo en la nube debido al uso de transformaciones espectrales y la complejidad del modelo ResNet50. Aquí viene parte de la necesidad de usar la versión Pro de Google Colab. Parte del tiempo asignado también se destinó a realizar una correcta división de los dos conjuntos de datos, con el objetivo de garantizar resultados coherentes y representativos.

8.2 Resultados visuales del entrenamiento

Se necesitaron 20 épocas en el entrenamiento del modelo. La Figura 8.1 presenta su evolución de la función de pérdida a lo largo del entrenamiento junto con el tiempo por época. Se observa una reducción significativa en las primeras iteraciones del modelo, estabilizándose progresivamente conforme va llegando a las últimas etapas.

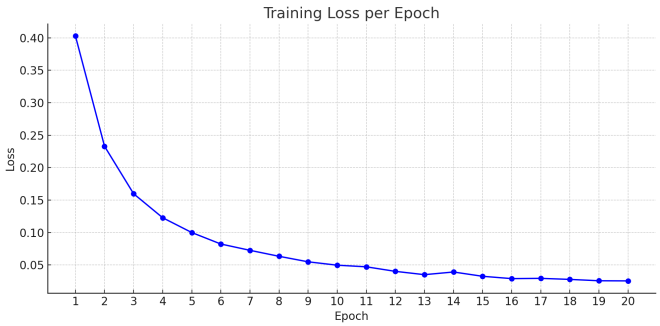

Figura 8.1: Evolución de la función de pérdida y el tiempo por época durante el entrenamiento.

8.3 Métricas de rendimiento

Para evaluar el rendimiento del sistema, se utilizaron métricas estándar en problemas de clasificación binaria. Estas se utilizaron sobre los conjuntos de validación conformado por 8,000 imágenes, con una distribución balanceada entre clases reales y falsas.

Accuracy: Se alcanzó una precisión del 97 %, lo que indica que el sistema acierta en 97 de cada 100 predicciones realizadas.

Precisión: La precisión indica qué proporción de las imágenes clasificadas como "falsasrealmente eran falsas, esto puede variar un poco con las pruebas que hice al correr el modelo pero el modelo tiene una mejor precisión con imágenes falsas.

- Fake: Precisión = 0.98
- Real: Precisión = 0.96

Recall (Sensibilidad): El recall indica qué proporción de las imágenes que verdaderamente son "falsas"fueron correctamente identificadas.

- Fake: Recall = 0.96
- Real: Recall = 0.98

F1-Score: Es la media armónica entre precisión y recall.

- Fake: F1-score = 0.97
- Real: F1-score = 0.97

Clase	Precisión	Recall	F1-score	Soporte
Real	0.96	0.98	0.97	3953
Fake	0.98	0.96	0.97	4047
Accuracy			0.97	8000
Macro promedio	0.97	0.97	0.97	8000
Promedio ponderado	0.97	0.97	0.97	8000

Tabla 8.1: Métricas de evaluación del modelo ResNEt50 para la clasificación de las imágenes reales y falsas.

Matriz de Confusión: La matriz de confusión permite evaluar el desempeño del modelo de clasificación binaria entre imágenes reales y falsas.

- **3879** casos fueron correctamente clasificados como **reales** (verdaderos negativos).
- **3865** casos fueron correctamente clasificados como **falsos** (verdaderos positivos).

- **74** imágenes reales fueron clasificadas erróneamente como falsas (falsos positivos).
- **182** imágenes falsas fueron clasificadas erróneamente como reales (falsos negativos).

$$\begin{bmatrix} 3879 & 74 \\ 182 & 3865 \end{bmatrix}$$

Esto muestra un rendimiento confiable del modelo, con una tasa muy baja de error tanto en falsos positivos como en falsos negativos, lo cual es crucial a la hora de elaborar tareas de detección de deepfakes donde las consecuencias de errores pueden ser significativas.

Otras métricas:

- ROC AUC Score: 0.9682
- MCC: 0.9364
- Balanced Accuracy: 0.9682

8.4 *Visualización de predicciones*

Se muestran algunos ejemplos de predicciones correctas realizadas por el modelo, destacando principalmente el espectro de Fourier en segundo plano. **Nota:** Intentamos adjuntar imágenes que causaron algún tipo de falsos positivos usando los datasets, pero nos fue muy difícil por la gran exactitud del modelo generado.

Figura 8.2: Imagen clasificada correctamente como Fake por el modelo.

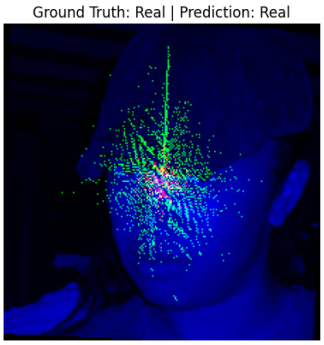

Figura 8.3: Imagen clasificada correctamente como Real por el modelo.

8.5 *Análisis visual extendido de predicciones*

Además de presentar métricas cuantitativas, se llevó a cabo una investigación cualitativa de análisis visual sobre un grupo de imágenes auténticas y otro de imágenes artificiales que no se emplearon en el entrenamiento o validación del modelo. Esta evaluación resulta importante y enriquece el trabajo porque permite observar de manera directa el comportamiento del modelo ante diversos tipos de datos, lo que facilita la interpretación visual de las decisiones que adopta durante el proceso clasificatorio. La Figura 8.4 ilustra algunos de estos ejemplos, en los cuales se ha aplicado el espectro de frecuencia (DFT) sobre los rostros identificados para facilitar el análisis de los patrones que el modelo podría estar empleando como fundamento para su predicción.

Cada recuadro exhibe el resultado del modelo en cada uno de los rostros; en determinadas circunstancias, algunos fueron recortados por YOLOv3. Muestra también su representación espectral y el resultado de la predicción encima de cada una de ellas. El título de cada imagen señala la etiqueta verdadera ("Ground Truth"), que muestra de qué dataset proviene, y la categorización realizada por el sistema ("Predicción"). Esto permite un contraste directo entre el rendimiento del modelo y la realidad. Al mirar estas imágenes, se muestra una tendencia de ciertas texturas o energías en el espectro de frecuencias que ayuda a diferenciar las imágenes reales de las que se crean de manera artificial. Aunque a simple vista no se puede observar un patrón claro, el modelo es excelente para clasificar cada una de ellas.

Esta ilustración visual mejora la comprensión del rendimiento del sistema híbrido, al mostrar cómo el modelo identifica diferencias sutiles en el dominio frecuencial, demostrando cómo el modelo logra identificar diferencias sutiles en el dominio frecuencial, que resultarían muy complicadas para las personas de distinguir. Refuerza además la hipótesis del estudio: que las transformaciones espectrales revelan patrones únicos que pueden ser aprovechables por redes como ResNet50 para mejorar la precisión en la detección de deepfakes.

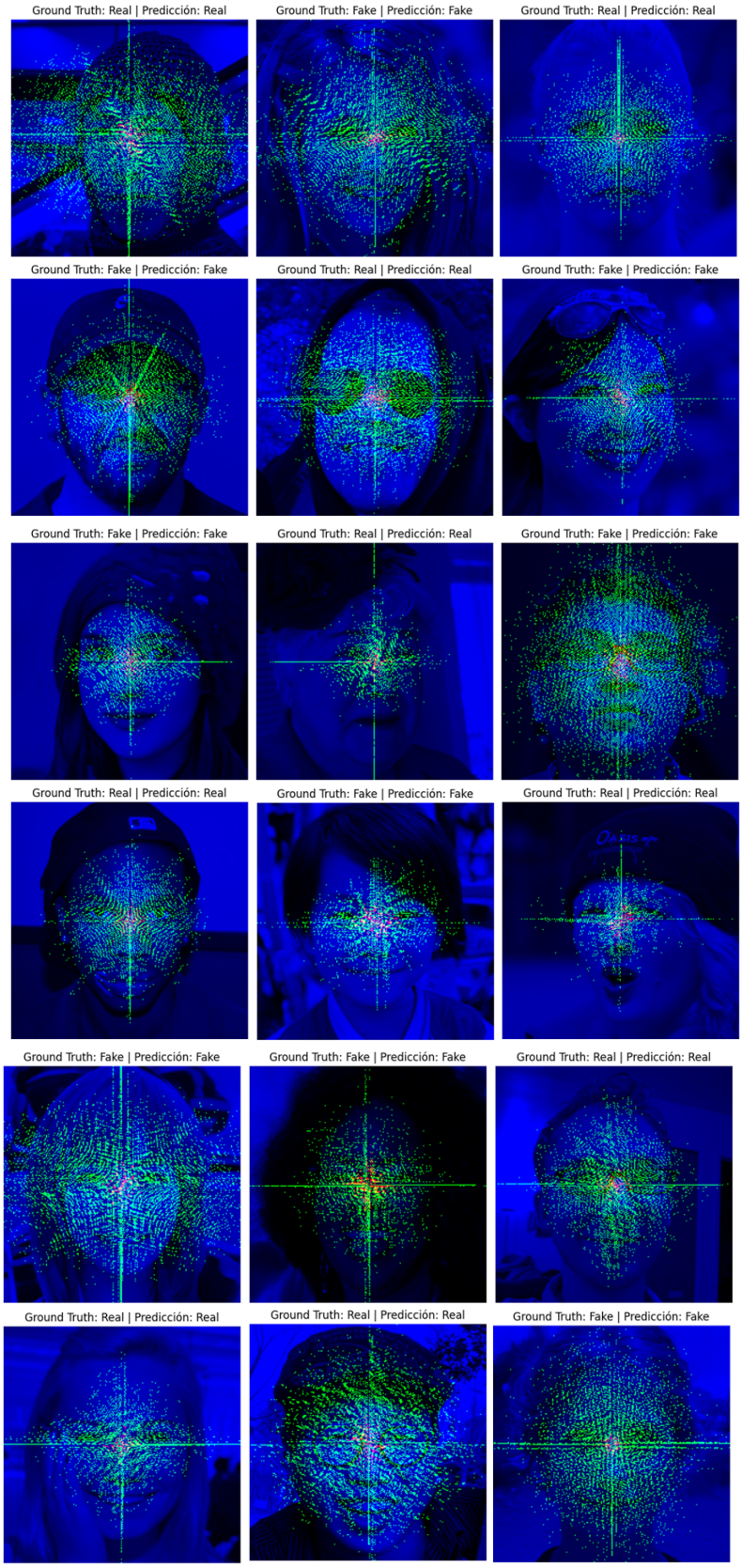

Figura 8.4: Ejemplos visuales extendidos: Predicciones del modelo con espectros DFT superpuestos sobre rostros detectados.

8.6 Comparación frente a métodos clásicos

Aunque no se realizó una comparación directa dentro de esta tesis con métodos clásicos, diversos estudios previos permiten establecer una base de referencia. En particular, enfoques tradicionales que emplean exclusivamente análisis espacial —como la detección de artefactos visuales, inconsistencias en la iluminación o el parpadeo artificial— han reportado precisiones que oscilan entre el 85 % y el 90 %. Por ejemplo, en el trabajo de investigación de Li et al. utilizaron patrones de movimiento ocular para detectar deepfakes con una exactitud cercana al 89 % [1], mientras que en la investigación de Matern et al. con una perspectiva de abordar el mismo problema, el principal énfasis fue la ausencia de parpadeo en los datasets generados por deepfake, logrando resultados muy similares [2].

El sistema propuesto en este trabajo, sumando todos los módulos que son la detección facial con la herramienta de YOLOv3 y análisis en el dominio frecuencial mediante la Transformada Discreta de Fourier (DFT), alcanzó una increíble precisión del 97 %. Este hallazgo indica que la utilización de representaciones espectrales proporciona datos extra que optimizan el proceso de clasificación y detección.

En contraste con los enfoques tradicionales, el modelo híbrido creado ofrece beneficios tanto en eficiencia como en solidez ante diversas formas de manipulación, lo que respalda la hipótesis de que el dominio frecuencial posee patrones pertinentes para reconocer contenido producido de manera artificial.

Este método que se propuso a lo largo del trabajo puede funcionar como un fundamento firme para sistemas automatizados de múltiples ámbitos que necesitan de la comprobación de autenticidad, particularmente en situaciones donde los métodos convencionales ya no son suficientes y se fundamentan exclusivamente en análisis visual, que tienen restricciones gracias a la evolución de las tecnologías que los generan.

[1] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. *arXiv preprint arXiv:1806.02877*, 2018. URL https://arxiv.org/abs/1806.02877

[2] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2019

8.7 Limitaciones observadas

- **Alto consumo computacional:** En este caso me vi en la necesidad de comprar y utilizar una herramienta externa para poder completar el entrenamiento y uso de mi modelo ya que ninguno de mis equipos locales podia consumir tantos recursos para el entrenamiento de la red y la division de los datos.

- **Prolongado tiempo de entrenamiento:** Es variable dependiendo de lo que se necesite, si se utilizan menos epocas en el trabajo y un conjunto de datos mas limitado puedes reducir el tiempo de

entrenamiento de la red, pero con el riesgo de perder acurracy y precisión en el modelo. En caso contrario tambien intente usar 120,000 imagenes del dataset y dure mas de un día intentado entrenar el modelo pero tuve que detenerlo porque estaba tardando demasiado hablando mas de 18 horas de continuo uso y encontrar ese balance de tiempo y trabajo.

- **Otros Datasets:** Aunque el modelo muestra un rendimiento sobresaliente con los datasets utilizados, ya se había anticipado que su desempeño podría no ser igual de efectivo al aplicarlo a otros conjuntos de datos que desconoce fuera de su entrenamiento. Esto sugiere la necesidad de entrenarlo con una mayor diversidad de herramientas de generación de deepfakes basadas en GANs, para que pueda generalizar mejor y adaptarse a distintas técnicas de falsificación.

8.8 *Repositorio del proyecto*

Con el objetivo de promover la transparencia, la reproducibilidad y el acceso abierto al conocimiento generado y su uso principalmente, se ha distribuido a mi nombre todo el código fuente del sistema híbrido sugerido en esta tesis en un repositorio público de GitHub.. Este repositorio incluye:

- Los scripts completos que fuero utilizados para la detección facial con la herramienta de YOLOv3.
- Los módulos de análisis espectral aplicando la Transformada Discreta de Fourier (DFT).
- El entrenamiento y la evaluación del clasificador ResNet-50.
- Notebooks de entrenamiento, visualización de resultados y pruebas experimentales.
- Instrucciones para ejecutar el pipeline completo de principio a fin.

El repositorio puede consultarse en la siguiente dirección: `https://github.com/Hatxu/Deepfake_Detection_with_Fourier`

Se alienta a las personas que estén leyendo este trabajo o a otros investigadores, estudiantes y personas de expertise interesados en el área a explorar, adaptar o mejorar este mismo trabajo como base para desarrollos futuros en la detección de deepfakes mediante análisis híbrido.

9 Conclusiones

Esta investigación tiene como objetivo resolver el problema de la identificación de imágenes producidas por herramientas de deepfake, sugiriendo un modelo híbrido que incorpora tres elementos que conforman el modelo: la identificación facial mediante el algoritmo de YOLOv3, el estudio en el dominio de la frecuencia mediante la Transformada Discreta de Fourier (DFT) y la clasificación final mediante una red neuronal convolucional ResNet-50.

Los hallazgos logrados durante el desarrollo del proyecto evidencian que la hipótesis planteada es factible. Esto se refiere al propósito de poder diferenciar imágenes auténticas de imágenes artificiales con alta exactitud en un entorno regulado. La aplicación de la transformada discreta de Fourier en el análisis posibilitó la identificación de patrones distintivos en imágenes falsas, resaltando la relevancia del análisis en el campo de la frecuencia como una posible alternativa o reemplazo a los métodos convencionales centrados exclusivamente en el análisis espacial.

Los resultados que se lograron en este proyecto dejan claro que el enfoque planteado funciona bien para diferenciar imágenes reales de las generadas artificialmente, especialmente en entornos controlados. El uso del dominio espectral ayudó a detectar ciertos patrones distintivos en imágenes falsas creadas con modelos como StyleGAN, lo que demuestra que analizar en frecuencia puede ser un muy buen complemento a los métodos tradicionales que solo miran lo espacial.

Lamentablemente, este sistema también posee ciertas limitaciones. Dentro de las más destacadas se incluyen su dependencia del tipo de dataset empleado, la ausencia de evaluación mediante videos o ambientes no regulados y la comparación de rendimiento con otras herramientas similares. Debido a las limitantes que tiene el modelo, se desarrolló el capítulo de Trabajo a Futuro, donde se proponen diversos caminos para la optimización y optimización del modelo.

En términos generales, este trabajo de obtención de grado busca ofrecer una contribución significativa en el campo de la detección de contenidos sintéticos, especialmente en un momento donde la tecnología de los deepfakes representa una amenaza creciente para

la veracidad de la información digital. El enfoque del modelo híbrido no solo demostró ser viable técnicamente, sino que también sienta las bases para soluciones más integrales que incorporen múltiples dominios de análisis que se pueden ir incorporando o mejorando en el pipeline, como en la parte espectral, la parte espacial y, en un futuro, la parte temporal.

9.1 *Aportaciones del trabajo*

Este trabajo presenta las siguientes contribuciones principales:

- Propuesta de un sistema híbrido que combina análisis espectral y visión computacional para la detección de deepfakes como enfoque principal.
- Implementación eficiente de una arquitectura basada en YOLOv3 para extracción facial automatizada y cómo se puede utilizar con el propósito de detección de deepfakes.
- Implementación de la Transformada Discreta de Fourier (DFT) para su aplicación en la conversión de los datasets públicos que se han manipulado previamente al campo de la frecuencia.
- Entrenamiento de la red neuronal ResNet-50 usando los campos espectrales obtenidos previamente y validación de estos con información no utilizada previamente para la tarea de clasificación de los resultados obtenidos.
- Análisis crítico de las limitaciones del sistema y propuesta de líneas de mejora realistas para su aplicación futura y una mejora continua.

9.2 *Implicaciones sociales y científicas*

El desarrollo de instrumentos para la detección de deepfakes continuamente ha evidenciado ser una necesidad en respuesta al aumento de temas como desinformación visual, la falsificación de identidad y la manipulación digital. Este esfuerzo se orienta hacia la creación de soluciones automáticas, éticas y técnicamente robustas que promuevan la detección de contenido sintético en medios digitales. Esta iniciativa tiene como objetivo preservar la confianza en la información visual y definir un futuro en el que sea esencial proseguir con esta línea de investigación.

9.3 *Reflexión*

Este proyecto supuso un enorme reto técnico, pero también me brindó un espacio para mi desarrollo personal, académico y profesional.

Poder incorporar los conocimientos obtenidos en la maestría me facilitó la creación de una respuesta auténtica a un problema en ascenso, fortaleciendo mi creencia de que la inteligencia artificial siempre debe dirigirse al beneficio social y moral. Por lo cual estaré siempre agradecido por la oportunidad de aportar con este trabajo.

9.4 *Publicación derivada del proyecto*

Parte del trabajo desarrollado en esta tesis ha sido estructurado y redactado como un artículo técnico bajo el título provisional *"ENHANCING DEEPFAKE DETECTION WITH YOLOV3 AND FOURIER TRANSFORM ANALYSIS"* a inicios de este mismo año. El artículo en sí fue evaluado para su posible publicación en la **IEEE International Conference on Image Processing (IEEE ICIP 2025)**, que se llevará a cabo del 14 al 17 de septiembre de 2025 en Anchorage, Alaska, Canadá. Esta propuesta surgió como mención de mi asesor de tesis, con la finalidad de organizar y confirmar las contribuciones de este trabajo en un entorno académico estricto y de alta especialización.

La IEEE ICIP es una conferencia reconocida como la más grande y prestigiosa, organizada por la Signal Processing Society, dedicada principalmente al procesamiento de visión por computadora, imágenes y video, temas que abarcan este trabajo de tesis. Este año, en su edición 2025, el evento tendrá como lema *"Imaging in the Age of GenAI"*, juntando a expertos, investigadores y profesionales de todas partes del mundo para discutir y mostrar avances técnicos de vanguardia y fomentar la colaboración interdisciplinaria. Más información sobre la conferencia puede consultarse en su sitio oficial: https://2025.ieeeicip.org.

El artículo enviado actualmente se encuentra bajo el proceso de revisión técnica, habiendo superado ya las etapas de verificación de derechos de autor y revisión inicial del documento. El contenido del trabajo resume los principales aportes técnicos mencionados en esta tesis, centrándose principalmente en el sistema híbrido expuesto que, como mencionamos durante todo este documento, integra la detección facial con YOLOv3, el análisis espectral mediante la Transformada Discreta de Fourier (DFT) y la clasificación a través de la red neuronal ResNet-50. La postulación a la conferencia IEEE ICIP 2025 constituye un avance crucial para la difusión internacional de este trabajo a la comunidad científica con fines de uso en las áreas de visión artificial, ciberseguridad e inteligencia artificial aplicada. Pienso que el hecho de que haya superado múltiples etapas de revisión dentro del proceso de publicación refuerza la validez del trabajo propuesto.

9.5 Cierre

Este trabajo señala el comienzo de una línea de investigación de la que deseo seguir siendo parte en mi carrera académica y laboral. Espero poder incluir nuevas técnicas y métodos que potencien el sistema que se ha planteado aquí, volviéndolo más sólido y flexible frente a los retos que plantea la evolución continua e inminente de estas tecnologías.

La detección de deepfakes es y seguirá siendo un desafío que se volverá más complicado conforme sigan avanzando los métodos y herramientas que generan esta clase de contenido, y mi objetivo es: Crear herramientas técnicas que faciliten su afrontamiento de forma eficaz y fiable. Aparte del valor científico que añade este trabajo, otro propósito al abordar esta investigación es que actúe como un llamado a la acción y a la reflexión. Adicionalmente, es una convocatoria para pensar acerca de los riesgos éticos, sociales y políticos que emergen en este nuevo siglo de manipulación digital. De igual manera, aspira a impulsar activamente la creación de nuevas herramientas que garanticen un mejor entorno digital y regulen sobre todo las herramientas existentes en el mercado, con el objetivo de crear un espacio digital seguro, transparente y confiable para nosotros y las generaciones venideras.

10 Trabajo a Futuro

Contenido

Este capítulo tiene como finalidad demostrar las posibles líneas de desarrollo y mejora del sistema híbrido propuesto en mi trabajo, así como también mencionar las limitaciones observadas y aspectos que no fueron contemplados o omitidos. Si bien los resultados obtenidos fueron satisfactorios en el entorno propuesto, también demuestran la viabilidad del enfoque implementado. Al igual que el avance acelerado de las técnicas de generación sintética, los nuevos desafíos tecnológicos hacen necesario proyectar el trabajo hacia futuras versiones más robustas, eficientes y aplicables, con enfoques más robustos que enriquezcan el trabajo propuesto.

10.1 Limitaciones y aspectos no contemplados del modelo

A pesar del desempeño satisfactorio del sistema híbrido expuesto en mi trabajo, es importante reconocer ciertas limitaciones y errores que no fueron abordados completamente en esta versión del modelo. Mencionar estos inconvenientes permite establecer una base para futuros trabajos y frutos de esta investigación y puedan abordarlos desde un principio desde su desarrollo.

- **Dependencia exclusiva de imágenes faciales:** El sistema fue diseñado para trabajar únicamente con rostros si bien esto también podría ser beneficio dependiendo en el contexto o el uso que se tenga en mente, también se limita por su aplicación a deepfakes que manipulan otras partes del cuerpo o que se centran en el entorno.
- **Sensibilidad al preprocesamiento:** El rendimiento del modelo depende en su mayoría la calidad del recorte facial generado por YOLOv3 [1] y en algunos casos en la forma en la que se presentan los datos, al tener que escalar o recortar partes de la imagen. Errores en esta etapa pueden introducir ruido o deformaciones que afectan negativamente la representación espectral.
- **Ausencia de análisis temporal:** Al centrarse únicamente en imágenes estáticas, el sistema no considera la continuidad temporal ni artefactos que emergen en secuencias de video[2].
- **Datos limitados y posible sesgo en el entrenamiento:** No se evaluaron variables demográficas o de contexto (iluminación, resolución, expresiones), lo que podria limitar la generalización del modelo. Este sesgo supone que basado en un dataset con demografia en especifico pueda limitar eficiencia en la clasificación con imagenes que no han sido tomadas en cuenta.
- **Suposición de espectros ideales:** El sistema parte del supuesto de que la Transformada Discreta de Fourier (DFT) permite distinguir con claridad entre imágenes reales y sintéticas. Sin embargo, al evaluar el modelo con conjuntos de datos que no formaban parte del entrenamiento, se evidenció una disminución significativa en su capacidad de generalización. Esto sugiere que, si bien el modelo ofrece un rendimiento sólido con imágenes similares a las vistas durante el entrenamiento, su eficacia se ve comprometida al enfrentarse a imágenes distintas en estilo, resolución o procedencia. Este comportamiento indica una alta dependencia del dominio de los datos y pone de manifiesto la necesidad de incorporar mecanismos de robustez y validación cruzada con datos más diversos[3].
- **Falta de resistencia a ataques adversarios:** No se evaluó la robustez del modelo ante manipulaciones diseñadas específicamente para

[1] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement, 2018. URL https://arxiv.org/abs/1804.02767. arXiv preprint arXiv:1804.02767

[2] David Guera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018

[3] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your step: Learning graphical representations for deepfake detection using frequency analysis. *arXiv preprint arXiv:2005.02791*, 2020

engañar al clasificador[4]. Si bien partimos de la fé del conjunto de datos, jamas tuvimos en cosideración ver que limitantes toma o no en cuenta el modelo.

- **Baja explicabilidad del modelo:** No se implementaron técnicas para interpretar o visualizar las razones por las que el modelo determina que una imagen es falsa o real[5]. Por lo que esta falta de información que podria ayudar a mejorar el modelo y indicar sus áreas de análisis.
- **Dependencia de los datasets:** El modelo mostró un buen rendimiento al trabajar con imágenes similares a las del conjunto de entrenamiento; sin embargo, al evaluarse con datasets distintos, su desempeño se vio considerablemente afectado. Esto revela una limitada capacidad de generalización y una fuerte dependencia del dominio de los datos. La herramienta, aunque efectiva dentro de un rango controlado, pierde precisión cuando se enfrenta a muestras provenientes de otros orígenes, resoluciones o condiciones de generación, lo cual limita su aplicabilidad en contextos abiertos o en escenarios reales no controlados.

[4] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2015

[5] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017

10.2 *Mejoras propuestas al sistema actual*

Existen diversas áreas que podrían optimizarse o ampliarse para incrementar el rendimiento, adaptabilidad y eficiencia del sistema. Algunas mejoras concretas incluyen:

- **Optimización del preprocesamiento facial:** Combinar YOLOv3 con técnicas de alineación o detección por puntos clave para mejorar la consistencia de los recortes.[6].
- **Conversión mejorada del espectro:** Explorar representaciones espectrales alternativas, como mapas de fase o espectros logarítmicos.[7].
- **Reducción del tiempo de inferencia:** Sustituir ResNet-50 por arquitecturas ligeras como EfficientNet[8] o MobileNet[9] para uso en tiempo real.
- **Optimización de recursos:** Mejorar la eficiencia en GPU y RAM mediante procesamiento por lotes y reducción de complejidad computacional.
- **Aumento de datos espectrales:** Aplicar técnicas de data augmentation que mantengan la integridad de las representaciones de frecuencia.
- **Evaluación avanzada:** Incorporar validación cruzada y nuevas métricas como AUC, curva ROC y sensibilidad-especificidad.

[6] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. In *IEEE Signal Processing Letters*, volume 23, pages 1499–1503, 2016

[7] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your step: Learning graphical representations for deepfake detection using frequency analysis. *arXiv preprint arXiv:2005.02791*, 2020

[8] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019

[9] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017

- **Interfaz gráfica del sistema:** Desarrollar una GUI que permita cargar imágenes, ver su espectro y recibir la clasificación final de forma accesible para usuarios no técnicos.

10.3 Líneas de investigación y aplicación futura

A continuación se presentan ideas para extender la investigación hacia nuevas áreas técnicas, científicas y sociales:

10.3.1 Extensión al análisis de video

El análisis de video permitiría capturar patrones temporales, artefactos de compresión y movimientos anómalos, mejorando la precisión de detección en contenido audiovisual[10].

10.3.2 Incorporación de audio y análisis multimodal

Integrar detección en el canal auditivo permitiría identificar inconsistencias entre imagen y voz. Un sistema multimodal sería más robusto ante nuevas formas de manipulación[11].

10.3.3 Actualización del sistema con nuevos generadores

Validar el sistema con deepfakes generados por tecnologías emergentes (como StyleGAN3, DALL·E, FaceFusion) permitiría mantenerlo actualizado y vigente frente a nuevos desafíos.

10.3.4 Exploración de nuevas arquitecturas de clasificación

Evaluar modelos como Vision Transformers (ViT)[12], ConvNeXt[13] o enfoques auto-supervisados permitiría mejorar la precisión, generalización y eficiencia del sistema.

10.3.5 Exploración de nuevos modelos de visión computacional o actualización a versiones más recientes

El modelo actual emplea YOLOv3 para la detección y recorte facial debido a su eficacia y amplia disponibilidad. No obstante, han surgido versiones más recientes como YOLOv5, YOLOv7 o YOLOv8[14], que presentan mejoras significativas en precisión, velocidad de inferencia y tamaño del modelo. La actualización a estas versiones podría optimizar el desempeño del sistema en escenarios más exigentes o en dispositivos con capacidades limitadas. Si bien, por temas éticos, no fueron considerados en el desarrollo de esta tesis, podrían ser explorados en un trabajo futuro.

[10] Essam Sabir, Jian Cheng, Abhinav Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9, 2019

[11] Tarun Mittal, Utsav Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2823–2832, 2020

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020

[13] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022

[14] Glenn Jocher et al. Yolo by ultralytics. https://github.com/ultralytics/yolov5, 2023. Accessed: 2024-03-20

Asimismo, podrían explorarse otras arquitecturas de visión por computadora como EfficientDet o modelos basados en Vision Transformers (ViT), que ofrecen un enfoque diferente al extraer representaciones más globales y jerárquicas de la imagen. Estas alternativas permitirían evaluar si una representación más abstracta de la región facial mejora la capacidad de distinguir entre imágenes reales y generadas artificialmente.

10.3.6 *Exploración de nuevos modelos de análisis espectral y ampliación al análisis espacial*

El análisis espectral basado en la Transformada Discreta de Fourier (DFT) ha sido una de las contribuciones clave de este trabajo. Sin embargo, también existen otras técnicas en el dominio de la frecuencia que podrían resultar más sensibles o informativas en la detección de artefactos sintéticos. Algunas de ellas incluyen la Transformada Wavelet Discreta (DWT), la Transformada Coseno Discreta (DCT) o incluso enfoques híbridos que combinan análisis multiescala y multifrecuencia.

Adicionalmente, se propone complementar el análisis en el dominio de la frecuencia con características extraídas en el dominio espacial. Esto permitiría al modelo beneficiarse tanto de patrones espectrales como de texturas visuales o artefactos geométricos presentes en las imágenes. Esta integración multimodal entre espectro y espacio podría mejorar la robustez del sistema, especialmente ante deepfakes más sofisticados que replican estructuras espectrales similares a las reales.

10.3.7 *Despliegue en plataformas prácticas*

Implementar el modelo en aplicaciones web, móviles o sistemas de vigilancia en tiempo real abriría el camino a soluciones prácticas con impacto directo en la sociedad.

10.3.8 *Explicabilidad y confianza del usuario*

Integrar mecanismos de interpretación como Grad-CAM, LIME o SHAP[15] permitiría visualizar las zonas espectrales más relevantes y aumentar la confianza del usuario en el modelo.

[15] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:4765–4774, 2017

10.3.9 *Refuerzo ético y normativo*

Colaboraciones con especialistas en ética digital y derecho podrían guiar el desarrollo de soluciones responsables, respetuosas de la privacidad y alineadas con normativas emergentes sobre IA[16].

[16] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1 (9):389–399, 2019

10.4 *Reflexión final*

En resumen, el sistema híbrido propuesto establece una base sólida para la detección de deepfakes, pero también revela múltiples oportunidades de mejora técnica, expansión funcional e integración interdisciplinaria. El avance de la inteligencia artificial generativa plantea retos constantes, y este trabajo busca no solo enfrentarlos desde una solución puntual, sino fomentar una visión crítica y responsable para su desarrollo futuro.

Bibliografía

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

Google AI, Joon Son Chung, and Andrew Zisserman. Deepfake detection challenge dataset (dfd). https://github.com/ondyari/FaceForensics, 2020. Accedido el 18 de abril de 2025.

Vito Nicola Convertini, Donato Impedovo, Ugo Lopez, Giuseppe Pirlo, and Gioacchino Sterlicchio. Discrete fourier transform in unmasking deepfake images: A comparative study of stylegan creations. *Information*, 15(711), 2024. DOI: 10.3390/info15110711. URL https://doi.org/10.3390/info15110711.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Ricard Durall, Margret Keuper, and Janis Keuper. Watch your step: Learning graphical representations for deepfake detection using frequency analysis. *arXiv preprint arXiv:2005.02791*, 2020.

Wired España. La huelga de actores de hollywood y la lucha contra la ia, 2023. URL https://shorturl.at/vAKKp.

European Commission. Proposal for a regulation on a european approach for artificial intelligence. Online, 2021. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2015.

Dario Guera and Edward J. Delp. Deepfake video detection using recurrent neural networks. *AVSS 2018 - IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2020. DOI: 10.1109/AVSS.2018.8639163.

David Guera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016a.

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 (CVPR)*, pages 770–778. IEEE, 2016b.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019.

Glenn Jocher et al. Yolo by ultralytics. https://github.com/ultralytics/yolov5, 2023. Accessed: 2024-03-20.

Tero Karras, Samuli Laine, and Timo Aila. Flickr-faces-hq dataset (ffhq). https://github.com/NVlabs/ffhq-dataset, 2019a. Accedido el 18 de abril de 2025.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019b.

Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. *arXiv preprint arXiv:1806.02877*, 2018. URL https://arxiv.org/abs/1806.02877.

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 *European conference on computer vision*, pages 740–755. Springer, 2014.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:4765–4774, 2017.

Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2019.

Tarun Mittal, Utsav Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2823–2832, 2020.

BBC Mundo. La crisis del porno deepfake que afecta a las escuelas coreanas, 2024. URL https://www.bbc.com/mundo/articles/c93p53292kyo.

Forbes México. Ia ha incrementando estafas con 'deepfakes', 2023. URL https://forbes.com.mx/ia-ha-incrementando-estafas-con-deepfakes.

Huy Nguyen, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *arXiv preprint arXiv:1906.06876*, 2019.

NumPy Developers. Numpy fft module documentation. https://numpy.org/doc/stable/reference/routines.fft.html, 2024. Último acceso: abril de 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 8024–8035, 2019.

Diálogo Político. La desinformación de la ia y las elecciones mundiales, 2025. URL https://dialogopolitico.org/edicion-especial-2025-democracia-artificial/la-desinformacion-de-la-ia.

Alexander Reben. 1 million fake faces dataset. https://archive.org/details/1mFakeFaces, 2019a. Accedido el 18 de abril de 2025.

Alexander Reben. 1 million fake faces. https://archive.org/details/1mFakeFaces, 2019b. Dataset generado con StyleGAN, licencia CC BY-NC 4.0.

Joseph Redmon. Tweet: I stopped doing computer vision research because... https://x.com/pjreddie/status/1230524770350817280, 2020. Accessed: 2025-04-18.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement, 2018. URL https://arxiv.org/abs/1804.02767. arXiv preprint arXiv:1804.02767.

Microsoft Research. Microsoft vision model resnet-50 combines web-scale data and multi-task learning to achieve state-of-the-art. https://shorturl.at/8MYeZ, 2023. Accessed: 2025-04-18.

Thomson Reuters. Deepfakes: Federal and state regulation, 2023. URL https://www.thomsonreuters.com/en-us/posts/government/deepfakes-federal-state-regulation.

Andreas Rössler, 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 (ICCV)*, pages 1–11, 2019.

Essam Sabir, Jian Cheng, Abhinav Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9, 2019.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020. DOI: 10.1016/j.inffus.2020.07.007.

Carnegie Mellon University. Deepfakes and the ethics of generative ai, 2023. URL https://tepperspectives.cmu.edu/all-articles/deepfakes-and-the-ethics-of-generative-ai.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. In *IEEE Signal Processing Letters*, volume 23, pages 1499–1503, 2016.