



EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA



TECNOLÓGICO
NACIONAL DE MÉXICO

Tecnológico Nacional de México

**Centro Nacional de Investigación
y Desarrollo Tecnológico**

Tesis de Maestría

**Modelo de diferenciación de rostros naturales y
producidos por GANs**

presentada por

Ing. Martín Ulises Millán Guerrero

como requisito para la obtención del grado de
Maestro en Ciencias de la Computación

Director de tesis

Dra. Andrea Magadán Salazar

Codirector de tesis

Dra. Daniela Alejandra Moctezuma Ochoa

Cuernavaca, Morelos, México. Septiembre de 2025.



Educación
Secretaría de Educación Pública



TECNOLÓGICO
NACIONAL DE MÉXICO



Centro Nacional de Investigación y Desarrollo tecnológico
Departamento de Ciencias Computacionales

Cuernavaca, Mor., 02/Septiembre/2025

OFICIO No. DCC/264/2025

Asunto: Aceptación de documento de tesis
CENIDET-AC-004-M14-OFICIO

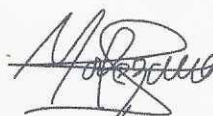
CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial de **Martín Ulises Millán Guerrero** con número de control **M23CE070**, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis de grado titulado **"Modelo de diferenciación de rostros naturales y producidos por GANs"** y hemos encontrado que se han atendido todas las observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

ATENTAMENTE

Excelencia en Educación Tecnológica®
"Conocimiento y Tecnología al Servicio de México"



Dra. Andrea Magadán Salazar
Directora de tesis

Dra. Daniela Alejandra Moctezuma Ochoa
Codirectora de tesis

Dr. Raúl Pinto Elías
Revisor 1

Dr. Nimrod González Franco
Revisor 2

C.c.p. Depto. Servicios Escolares.
Expediente / Estudiante



2025
Año de
La Mujer

Interior Internado Palmira S/N, Col. Palmira,

cenidet
Centro Nacional de Investigación
y Desarrollo Tecnológico

C. P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 3201,

e-mail: dcc.cenidet@ecnmx tecnmx | cenidet.tecnmx





Centro Nacional de Investigación y Desarrollo tecnológico
Subdirección Académica

Cuernavaca Mor, 06/octubre/2025

Oficio No. SAC/235/2025

Asunto: Autorización de impresión de tesis

MARTÍN ULISES MILLÁN GUERRERO
CANDIDATO AL GRADO DE MAESTRO
EN CIENCIAS DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"Modelo de diferenciación de rostros naturales y producidos por GANs"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

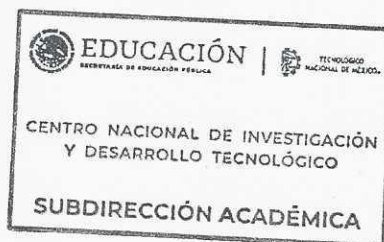
Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

ATENTAMENTE

Excelencia en Educación Tecnológica®

"Conocimiento y Tecnología al Servicio de México"

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO



c.c.p. Departamento de Ciencias Computacionales
Departamento de Servicios Escolares

CMAZ/lmz



2025
Año de
La Mujer
Indígena

Interior Internado Palmira S/N, Col. Palmira,
C. P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 4104,
e-mail: acad_cenidet@tecnm.mx tecnm.mx | cenidet.tecnm.mx

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico



Dedicatoria

A mi abuelita Tere,
por enseñarme lo que es el amor.

Agradecimientos

Me gustaría manifestar mi agradecimiento a la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (*SECIHTI*) por el respaldo financiero que me brindaron para llevar a cabo mis estudios de maestría. Este respaldo resultó crucial para realizar este estudio.

Además, mi agradecimiento se dirige al Centro Nacional de Investigación y Desarrollo Tecnológico (*TecNM/CENIDET*), por proporcionarme el entorno y los medios indispensables que posibilitaron la realización de este proyecto de tesis.

A la Dra. Andrea Magadán Salazar, mi directora de tesis, y a la Dra. Daniela Alejandra Moctezuma Ochoa, mi codirectora, les agradezco profundamente su orientación y consejo durante todo el proceso de elaboración de este trabajo. Su compromiso, recomendaciones, respaldo, tiempo y paciencia resultaron ser inestimables para finalizar este trabajo.

Mi agradecimiento a mis revisores de tesis, el Dr. Raúl Pinto Elías y el Dr. Nimrod González Franco, por sus observaciones y evaluaciones, que resultaron esenciales para la elaboración de esta tesis.

De forma particular, deseo expresar mi gratitud a mis padres, Benito y Gabriela, por su respaldo incondicional y continuo a lo largo de toda mi existencia. Su amor y aliento fueron mi mayor incentivo. Además, a mis hermanos, por su constante presencia.

Y a mi novia Ailen, por estar siempre a mi lado, gracias a tu paciencia y tus ánimos, sin ti todo esto no habría sido lo mismo. Gracias por acompañarme en cada paso de esta aventura.

Resumen

Esta tesis aborda el problema de la creciente dificultad para distinguir entre rostros naturales y aquellos generados artificialmente mediante inteligencia artificial. La proliferación de imágenes falsas y la manipulación de información en línea han hecho crucial el desarrollo de un sistema de reconocimiento preciso. El objetivo general de este proyecto es, por lo tanto, desarrollar un modelo que identifique si un rostro en una imagen fue generado artificialmente o no, mediante el análisis de imágenes y *machine learning*.

La metodología empleada incluyó el estudio del estado del arte para comprender cómo se caracterizan los rostros naturales y artificiales, así como el análisis de herramientas y redes generadoras de rostros artificiales. Se recopilaron conjuntos de datos de rostros reales y sintéticos. Se implementó una red neuronal para el reconocimiento de rostros y se evaluó cuantitativamente el sistema desarrollado, utilizando métricas de rendimiento de modelos de *machine learning*. Además, se utilizaron técnicas de aumento de datos, incluyendo *Deep Dream* y deformación de imágenes. La experimentación involucró más de 10 arquitecturas, de las cuales se seleccionó la mejor (MViTv2), a la cual se le realizaron modificaciones para mejorar su capacidad de generalización.

Los resultados más importantes incluyen la obtención de un modelo con una exactitud promedio del 98.39% en la clasificación de rostros. Además, se identificaron áreas de mejora, como la robustez frente a modelos de difusión y técnicas de post-procesamiento. La inclusión de técnicas de aumento de datos como *Deep Dream* y deformación de imágenes fue crucial para mejorar el rendimiento del modelo.

En conclusión, se desarrolló un modelo eficaz para la detección de rostros generados artificialmente, con áreas de mejora identificadas para futuras investigaciones. El modelo final cumplió con el objetivo general del proyecto, proporcionando una herramienta robusta para discernir entre rostros naturales y sintéticos, en diversos escenarios.

Abstract

This thesis addresses the problem of the growing difficulty in distinguishing between natural faces and those artificially generated using artificial intelligence. The proliferation of fake images and the manipulation of online information have made the development of an accurate recognition system crucial. The overall objective of this project is, therefore, to develop a model that identifies whether a face in an image was artificially generated or not, through the analysis of images and machine learning.

The methodology employed included studying the state of the art to understand how natural and artificial faces are characterized, as well as the analysis of tools and generative networks for artificial faces. Datasets of both real and synthetic faces were collected. A neural network was implemented for face recognition and the developed system was quantitatively evaluated, using machine learning model performance metrics. Additionally, data augmentation techniques were used, including Deep Dream and image deformation. The experimentation involved more than 10 architectures, from which the best one (MViTv2) was selected, and modifications were made to it to improve its generalization capability.

The most important results include obtaining a model with an average accuracy of 98.39% in face classification. Furthermore, areas for improvement were identified, such as robustness against diffusion models and post-processing techniques. The inclusion of data augmentation techniques like Deep Dream and image deformation was crucial for improving the model's performance.

In conclusion, an effective model for the detection of artificially generated faces was developed, with areas for improvement identified for future research. The final model met the overall objective of the project, providing a robust tool to discern between natural and synthetic faces in various scenarios.

Contenido

I. Capítulo I: Planteamiento del problema	1
1.1. Introducción.....	1
1.2. Problema.....	3
1.3. Objetivos	3
1.4. Alcances y limitaciones.....	4
1.5. Metodología	5
1.6. Organización de la tesis.....	8
II. Capítulo II: Marco teórico	10
2.1. Fundamentos y tecnologías base.....	10
2.2. Generación y manipulación de imágenes.....	14
2.3. Métricas de rendimiento.....	21
2.4. Modelos de transformadores visuales	23
III. Capítulo III: Estado del arte	27
3.1. Trabajos relacionados.....	27
3.2. Estado del arte	28
IV. Capítulo IV: Conjuntos de datos.....	36
4.1. Conjuntos de datos de rostros naturales	36
4.2. Conjuntos de datos de rostros generados.....	40
V. Capítulo V: Desarrollo del sistema	47
5.1. Entorno de desarrollo.....	47
5.2. Experimentación	48
VI. Capítulo VI: Resultados.....	69
6.1. Resumen del modelo final	69
6.2. Evaluación de la eficacia de los métodos de aumento de datos propuestos, en el ámbito de clasificación de rostros generados artificialmente	71
6.3. Escenarios de evaluación	73
6.4. Análisis de los resultados	82
VII. Capítulo VII: Conclusiones	84
7.1. Objetivos logrados	85
7.2. Contribuciones	87
7.3. Trabajos futuros	87
7.4. Actividades académicas	89
VIII. Referencias.....	93

Índice de figuras

Figura 1.1. Perfil falso de LinkedIn con fotografía sintetizada artificialmente [1].	2
Figura 1.2. Primer renglón: rostros artificiales [2]. Segundo renglón: rostros naturales [3].....	2
Figura 1.3. Metodología de desarrollo.	5
Figura 1.4 Metodología final.....	8
Figura 1.5. Diagrama de inferencia.....	8
Figura 2.1. Etapas fundamentales del tratamiento digital de imágenes [7].....	11
Figura 2.2. Categorización de falsificaciones profundas visuales, adaptada de [13].	15
Figura 2.3. Mejora en la calidad de los rostros generados por variaciones de GANs [13].....	17
Figura 2.4. Rarezas en la imagen sintetizada [1].	17
Figura 2.5 Vista general del modelo ViT [28].....	24
Figura 2.6 Red piramidal de características [29].	25
Figura 2.7 Configuración de las variantes de MVITv2 [29].	26
Figura 2.8 Efectos del preentrenamiento con ImageNet en los modelos [29].....	26
Figura 4.1. Imágenes obtenidas del conjunto CelebA-HQ [95].	37
Figura 4.2. Imagen obtenida del conjunto FFHQ [3].....	38
Figura 4.3. Ejemplo del conjunto LSUN [99].....	39
Figura 4.4. Ejemplos del conjunto AFHQ [100].....	39
Figura 4.5. Imagen obtenida del conjunto ProGAN [95].	41
Figura 4.6. Imagen obtenida del conjunto StyleGAN [3].....	42
Figura 4.7. Artefactos en forma de gota en imágenes generadas por StyleGAN [2].	43
Figura 4.8. Imagen obtenida del conjunto StyleGAN2 [2].....	43
Figura 4.9. Imagen obtenida del conjunto StyleGAN2-ADA [101].	44
Figura 4.10. Imágenes obtenidas del conjunto StyleGAN3. StyleGAN3-T (superior izquierda). StyleGAN3-T (superior derecha). StyleGAN3-U-R (inferior izquierda). StyleGAN3-U-T (inferior derecha) [93].....	45
Figura 5.1 Gráficas de entrenamiento de la experimentación inicial.....	51
Figura 5.2 Gráficas de entrenamiento de la experimentación con aumento de datos.	53
Figura 5.3 Gráficas de entrenamiento utilizando el modelo 'vit_tiny_patch16_224'.	54
Figura 5.4 Gráficas de entrenamiento con el conjunto de 5,000 muestras.	55
Figura 5.5 Comparación de optimizadores en tareas de clasificación [111].	58
Figura 5.6 Gráficas de entrenamiento al aplicar cambios en el dropout, optimizador y ajuste de tasa de aprendizaje.	59
Figura 5.7 Gráficas de entrenamiento al implementar épocas de calentamiento. .	60
Figura 5.8 Gráficas de entrenamiento al utilizar 'Focal Loss' como función de pérdida.....	61
Figura 5.9 Gráficas de entrenamiento al congelar los bloques 0, 1 y 2 de la arquitectura.....	61

Figura 5.10 Gráficas de entrenamiento después de ajustar el tamaño de entrada del modelo.....	63
Figura 5.11 Ejemplos de imágenes con acercamiento en artefactos.....	64
Figura 5.12 Ejemplos de representaciones internas de los modelos [93].....	65
Figura 5.13 Cambios en la imagen al aplicar filtro Deep Dream en una zona.	66
Figura 5.14 Ejemplos de imágenes generadas artificialmente con deformaciones.	66
Figura 5.15 Cambios en la imagen al aplicar deformación en una zona.....	67
Figura 6.1. Resultados del escenario 4.	79
Figura 6.2. Ejemplos de rostros generados con el modelo de difusión.....	80
Figura 6.3. Resultados del escenario 5.	80
Figura 6.4. Izquierda: Imágenes originales. Centro: Imágenes a 64x64. Derecha: Imágenes restauradas.....	81
Figura 6.5. Resultados del escenario 6.	82
Figura 7.1. Evidencia y certificado de participación.....	89
Figura 7.2. Certificado de finalización de estancia.	90
Figura 7.3. Certificado de exposición.....	90
Figura 7.4. Reconocimiento obtenido en ponencia.	91
Figura 7.5. Certificado obtenido.	92
Figura 7.6. Certificado obtenido.	92

Índice de tablas

Tabla 2.1. Tipos de falsificaciones profundas reportadas, adaptada de [13].	16
Tabla 2.2. Matriz de confusión para dos clases. Las celdas indican el número de positivos verdaderos (TP), falsos positivos (FP), negativos verdaderos (TN) y falsos negativos (FN) [24].	21
<i>Tabla 3.1. Síntesis tabular de los trabajos examinados relacionados al reconocimiento de rostros generados.</i>	29
<i>Tabla 3.2. Síntesis tabular de los artículos de revisión examinados.</i>	31
Tabla 3.3. Herramientas populares generadoras de rostros.	31
Tabla 4.1. Resumen de los conjuntos de datos.	46
Tabla 5.1. Modulos de Python utilizados.	48
Tabla 5.2. Conjunto de datos de entrenamiento y de la validación.	49
Tabla 5.3 Conjunto de datos para las pruebas.	50
Tabla 5.4. Exactitud (%) en las pruebas posteriores de la experimentación inicial.	51
Tabla 5.5. Técnicas de aumento de datos utilizadas.	52
Tabla 5.6. Exactitud (%) en las pruebas posteriores de la experimentación con aumento de datos.	52
Tabla 5.7. Exactitud (%) en las pruebas posteriores utilizando el modelo 'vit_tiny_patch16_224'.	53
Tabla 5.8. Conjunto de datos de entrenamiento y de la validación con 5,000 muestras.	54
Tabla 5.9. Exactitud (%) en las pruebas posteriores con el conjunto de 5,000 muestras.	54
Tabla 5.10. Exactitud (%) en los modelos utilizados.	56
Tabla 5.11. Exactitud (%) en los modelos MViTv2.	57
Tabla 5.12. Exactitud (%) en las pruebas posteriores al aplicar cambios en el dropout, optimizador y ajuste de tasa de aprendizaje.	59
Tabla 5.13. Exactitud (%) en las pruebas posteriores al implementar épocas de calentamiento.	60
Tabla 5.14. Exactitud (%) en las pruebas posteriores al utilizar 'Focal Loss' como función de pérdida.	60
Tabla 5.15. Exactitud (%) en las pruebas posteriores al congelar los bloques 0, 1 y 2 de la arquitectura.	62
Tabla 5.16. Exactitud (%) en las pruebas posteriores al congelar únicamente el bloque 0 de la arquitectura.	62
Tabla 5.17 Modificaciones en el conjunto de entrenamiento.	62
Tabla 5.18. Exactitud (%) en las pruebas posteriores al congelar únicamente el bloque 0 de la arquitectura.	63
Tabla 5.19. Exactitud (%) en las pruebas posteriores después de ajustar el tamaño de entrada del modelo.	63
Tabla 5.20 Modificación del conjunto de prueba.	68
Tabla 5.21. Conjunto de entrenamiento al realizar ajuste del volumen de aumento de datos.	68
Tabla 6.1. Conjunto de datos de entrenamiento, incluyendo el aumento de datos fuera de línea.	69

Tabla 6.2. Técnicas de aumento de datos utilizadas en línea.....	70
Tabla 6.3. Parámetros del entrenamiento de la red.	70
Tabla 6.4. Resultados en términos de exactitud con las configuraciones probadas.	72
Tabla 6.5. Métricas del modelo final.	73
Tabla 6.6. Exactitud del modelo en los conjuntos probados. * Significa que el conjunto está presente en el entrenamiento.....	73
Tabla 6.7. Comparación con trabajos del estado del arte en términos de exactitud.	74
Tabla 6.8. Matriz de confusión del modelo.	74
Tabla 6.9. Resultados de exactitud en pruebas con imágenes con procesamiento posterior a la generación.....	75
Tabla 6.10. Comparación en términos de TPR/TNR con el estado del arte.	76
Tabla 6.11. Comparación en términos de exactitud con el estado del arte.	76
Tabla 6.12. Exactitud del modelo en los conjuntos probados; con imágenes que no contenían rostros.....	78
Tabla 7.1. Objetivos específicos cumplidos.....	86

Acrónimos

Acrónimo	Significado en inglés	Significado en español
ACC	Accuracy	Exactitud
AdaIN	Adaptive Instance Normalization	Normalización adaptativa de instancias
ADA	Adaptive Discriminator Augmentation	Aumento del discriminador adaptativo
AFHQ	Animal Faces High Quality	Rostros de animales de alta calidad
CELEBA	CelebFaces Attributes Dataset	Conjunto de datos de atributos de rostros de celebridades
CNN	Convolutional Neural Network	Red neuronal convolucional
DD	Deep Dream	Sueño profundo
EG3D	Efficient Geometry-aware 3D	3D eficiente consciente de la geometría
FFHQ	Flicker Faces High Quality	Caras <i>Flicker</i> de alta calidad
FID	Frechet Inception Distance	Distancia Frechet Inception
FN	False Negative	Falso negativo
FP	False Positive	Falso positivo
FPN	Feature Pyramid Network	Red de pirámide de características
FPR	False Positive Rate	Tasa de falsos positivos
GAN	Generative Adversarial Network	Red generativa antagónica
GFPGAN	Generative Facial Prior Generative Adversarial Network	Red generativa antagónica con prior facial generativo
LSUN	Large-scale Scene Understanding	Comprensión de escenas a gran escala
MVIT	Multiscale Vision Transformers	Transformadores de visión multiescala
PROGAN	Progressive Growing of Generative Adversarial Networks	Crecimiento progresivo de redes generativas antagónicas
ROC	Receiver Operating Characteristic	Característica operativa del receptor
SGD	Stochastic Gradient Descent	Descenso de gradiente estocástico
TN	True Negative	Verdadero negativo
TP	True Positive	Verdadero positivo
TPR	True Positive Rate	Tasa de verdaderos positivos

Acrónimos (continuación)

Acrónimo	Significado en inglés	Significado en español
VAE	Variational Auto Encoder	Codificador automático variacional.
ViT	Visual Transformer	Transformador visual
W	Warping	Deformación

Glosario

Término	Definición
Ajuste Fino	Proceso de tomar un modelo de red neuronal que ya ha sido entrenado en un gran conjunto de datos y re-entrenarlo en un conjunto de datos más pequeño y específico.
Algoritmo de optimización	Motor que ajusta los parámetros internos de una red neuronal durante el entrenamiento. Su objetivo es minimizar la "función de pérdida", es decir, reducir el error del modelo.
Aprendizaje por transferencia	Técnica de aprendizaje automático en la que un modelo entrenado para una tarea se reutiliza como punto de partida para una segunda tarea relacionada.
Aumento de datos	Crear artificialmente nuevas muestras de datos para el entrenamiento a partir de las existentes, aplicando transformaciones como rotaciones, cambios de color o recortes.
CrossEntropy	Una función de pérdida comúnmente utilizada en problemas de clasificación para medir qué tan acertadas son las predicciones del modelo en comparación con las etiquetas reales.
Deep Dream	Una técnica de aumento de datos para simular artefactos de imágenes generadas.
Deepfake	Contenido multimedia (imágenes o videos) manipulado mediante técnicas de aprendizaje profundo para crear falsificaciones realistas, como intercambiar el rostro de una persona por el de otra.
Deformación de imágenes (Image Warping)	Técnica que modifica digitalmente la geometría de una imagen, alterando la posición de los píxeles para crear distorsiones.

Glosario (continuación)

Término	Definición
Dropout	Una técnica de regularización que "apaga" neuronas al azar durante el entrenamiento de una red neuronal. Esto obliga al modelo a aprender características más robustas y evita que dependa en exceso de neuronas específicas, reduciendo así el sobreajuste.
Épocas de calentamiento	Un período inicial durante el entrenamiento de un modelo en el que se utiliza una tasa de aprendizaje muy baja.
Focal loss	Una función de pérdida diseñada para problemas de clasificación con un desequilibrio entre clases.
Función de pérdida	Una fórmula matemática que calcula el error entre las predicciones de un modelo y los valores reales.
Sobreajuste	Un problema que ocurre cuando un modelo de aprendizaje automático se ajusta demasiado bien a los datos de entrenamiento y, como resultado, pierde su capacidad de generalizar y hacer predicciones precisas sobre datos nuevos que no ha visto antes.
Tasa de aprendizaje	Parámetro que controla la magnitud de los ajustes que el optimizador realiza en los pesos del modelo durante el entrenamiento.
Transformador visual	arquitectura de red neuronal que aplica el concepto de "transformadores", originalmente desarrollado para el procesamiento del lenguaje, a tareas de visión por computadora. En lugar de procesar píxeles, divide la imagen en parches y analiza las relaciones entre ellos.

Capítulo I: Planteamiento del problema

1.1. Introducción

En la era digital actual, la creación de rostros falsos mediante herramientas de generación de imágenes ha alcanzado un punto alarmante. En un mundo donde la inteligencia artificial puede generar retratos casi indistinguibles de las personas reales, surgen preocupaciones significativas sobre la autenticidad y la integridad de las imágenes en línea.

El problema central al que se enfrenta radica en la necesidad de distinguir con precisión entre los rostros naturales (aquellos que pertenecen a individuos reales) y los rostros generados artificialmente por medio de tecnología de inteligencia artificial. Esta distinción se ha vuelto crucial en un mundo donde la proliferación de imágenes falsas y la manipulación engañosa de la información son comunes.

Para comprender la magnitud del problema, es crucial notar cómo el aumento de imágenes generadas por inteligencia artificial ha creado una situación preocupante. Estas imágenes, utilizadas con fines maliciosos, incluyen desde la creación de contenido engañoso hasta la manipulación de identidades y la vulnerabilidad de sistemas de seguridad.

La Figura 1.1 muestra un ejemplo de un perfil falso [1] en la plataforma *LinkedIn*, en el que se utiliza una imagen generada artificialmente. Este perfil ficticio es un ejemplo palpable de cómo las imágenes generadas mediante inteligencia artificial son utilizadas de manera engañosa en plataformas de redes sociales y perfiles en línea.

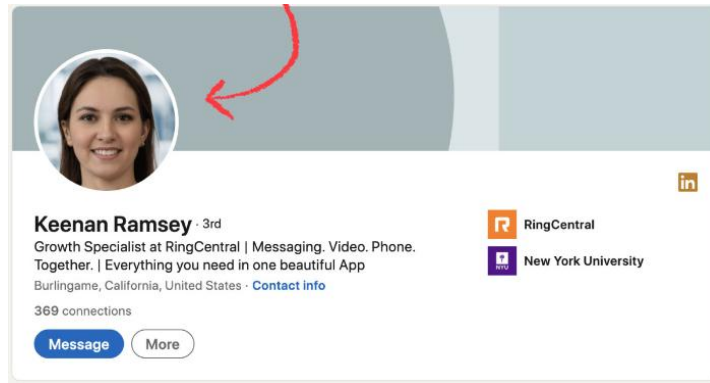


Figura 1.1. Perfil falso de LinkedIn con fotografía sintetizada artificialmente [1].

Por tanto, en respuesta a esta creciente preocupación, es fundamental avanzar en el desarrollo de un sistema de reconocimiento capaz de diferenciar de manera efectiva entre rostros naturales y aquellos generados artificialmente. Este sistema no solo actuará como un escudo esencial contra la manipulación y la desinformación en el ámbito digital, sino que también desempeña un papel crucial en la preservación de la integridad en la representación de individuos en imágenes.

Además de las preocupaciones mencionadas, otro aspecto relevante a considerar es la confianza del usuario en los medios visuales. En un entorno donde las imágenes pueden manipularse con relativa facilidad, hay un riesgo de que los usuarios cuestionen la veracidad de la información visual en línea. Esta pérdida de confianza en los medios visuales puede tener efectos perjudiciales, disminuyendo la credibilidad de la información y disminuyendo la eficacia de la comunicación en un mundo donde la imagen desempeña un papel cada vez más predominante.

En la Figura 1.2, se presentan ejemplos de rostros naturales y generados artificialmente con el propósito de subrayar lo complejo que puede resultar, a simple vista, distinguir entre ambos tipos de imágenes. Esto ilustra la magnitud del desafío en la era actual de la manipulación de imágenes generadas artificialmente.



Figura 1.2. Primer renglón: rostros artificiales [2]. Segundo renglón: rostros naturales [3].

La importancia de abordar este problema se vuelve aún más evidente en un entorno donde la tecnología avanza rápidamente y donde la confiabilidad de las imágenes visuales es fundamental para garantizar la seguridad y la autenticidad en línea. La tarea de distinguir entre lo real y lo artificial en el ámbito de los rostros humanos se ha vuelto un desafío crucial, cuyo abordaje no solo contribuirá a la preservación de la verdad y la autenticidad, sino que también fortalecerá la capacidad para enfrentar las amenazas digitales en una era marcada por la proliferación de información visual falsa.

1.2. Problema

En la era actual de avances tecnológicos y desarrollos en inteligencia artificial, la generación de rostros humanos sintéticos mediante herramientas de procesamiento de imágenes se ha convertido en una realidad asombrosa. Estas aplicaciones de generación de rostros tienen la capacidad de producir imágenes fotorrealistas de personas que no existen en la realidad. Esto plantea preocupaciones y desafíos importantes en varios contextos, incluyendo la verificación de identidad y la seguridad en general. La capacidad de crear y difundir rostros falsos de manera convincente puede tener implicaciones significativas en la propagación de información engañosa, la manipulación de identidades y la vulnerabilidad de sistemas de seguridad.

El problema central que la tesis aborda es desarrollar un sistema para detectar diferencias entre rostros naturales, es decir, aquellos pertenecientes a individuos reales y rostros generados artificialmente por medio de algoritmos de inteligencia artificial.

1.3. Objetivos

Objetivo general

Desarrollar un modelo que identifique si el rostro en una imagen es generado artificialmente o no, mediante el análisis de imágenes.

Objetivos específicos

- Estudiar los diferentes enfoques que permiten la caracterización de un rostro natural y uno artificial, mediante el análisis del estado del arte.
- Analizar herramientas y redes generadoras de rostros artificialmente.
- Contar con conjuntos de rostros creados con herramientas de Inteligencia Artificial y rostros reales.

- Implementar una red neuronal para el reconocimiento entre rostros naturales y aquellos creados mediante herramientas de inteligencia artificial.
- Evaluar el rendimiento del sistema desarrollado de manera cuantitativa a través de las métricas tradicionales en el área de reconocimiento de patrones.

1.4. Alcances y limitaciones

Alcances

- El sistema es capaz de identificar rostros naturales y sintéticos (creados por herramientas de generación de rostros) incluso en condiciones no controladas.
- El sistema analiza y evalúa las imágenes creadas de al menos 2 arquitecturas de generación de imágenes de rostros.
- El sistema de reconocimiento de rostros permite la adaptación a diferentes escalas, perspectivas y fondos.

Limitaciones

- El sistema desarrollado no proporciona información relacionada a la identidad de la persona, género, edad ni emoción presente.
- El sistema analiza exclusivamente imágenes que contengan un solo rostro.
- El sistema únicamente analiza imágenes estáticas.
- No se garantiza un funcionamiento efectivo del sistema en casos de imágenes de muy baja calidad o cuando se trata de escaneos o reproducciones de imágenes.
- El sistema no garantiza el correcto funcionamiento si existe oclusiones en el rostro como lentes de sol y cubrebocas.
- El sistema desarrollado no aborda la detección de intercambio de rostros, sincronización labial, control de títeres o manipulación de atributos faciales.
- El sistema se diseña para trabajar exclusivamente con imágenes de individuos con características faciales típicas y saludables, excluyendo la consideración de variaciones relacionadas con condiciones médicas, diferencias notables en la apariencia facial o anomalías en la estructura ocular.
- El sistema no considera rostros en condiciones de iluminación extremadamente baja o alta.

1.5. Metodología

Para abordar el problema de la diferenciación de rostros, la propuesta de solución se centra en el desarrollo de un modelo de clasificación basado en un transformador visual (*MViTv2*), utilizando un ajuste de parámetros y, adicionalmente, enriquecido con técnicas de aumento de datos no convencionales, como *Deep Dream* y deformación de imágenes. Este enfoque busca entrenar al modelo no solo para reconocer rostros, sino para identificar los artefactos sutiles y las texturas anómalas características de las imágenes generadas artificialmente. La Figura 1.3 ilustra los puntos principales de este proceso. La metodología cuenta de las siguientes etapas clave, detalladas a lo largo del Capítulo 5:

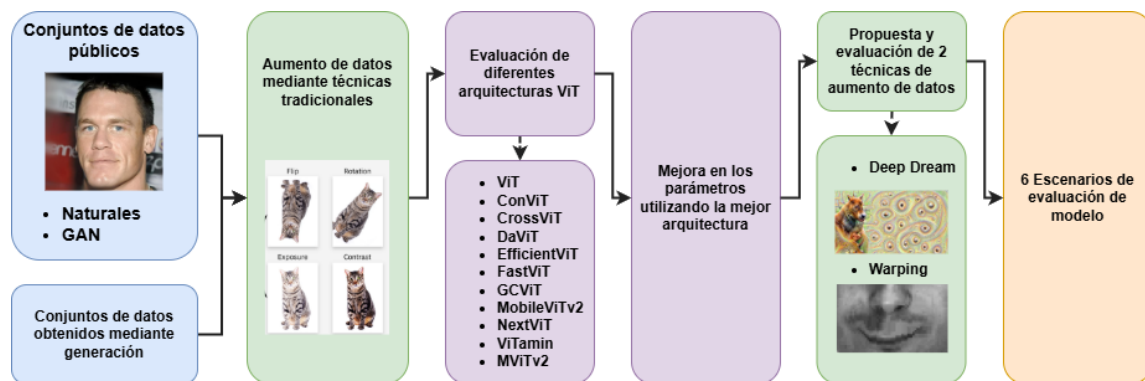


Figura 1.3. Metodología de desarrollo.

1. Para contar con ejemplos de rostros diversos y representativos del problema se integró el conjunto de datos compuesto por dos clases: rostros naturales y rostros generados mediante GANs. Para la clase de rostros naturales, se utilizaron los conjuntos públicos como son *FFHQ* y *CelebA-HQ*. Para la clase de rostros generados, se emplearon conjuntos creados por arquitecturas GAN como *ProGAN*, *StyleGAN*, *StyleGAN2*, *StyleGAN2-ADA* y *StyleGAN3* que son los algoritmos base que utilizan las actuales herramientas generadoras de rostros.
2. Además, se aplicaron técnicas de aumento de datos en línea tradicionales, incluyendo transformaciones geométricas, aumento del espacio de color, filtrado de kernel, borrado aleatorio, entre otras, para incrementar la diversidad del conjunto de entrenamiento al considerar otros posibles escenarios en las imágenes y mejorar la capacidad de generalización del modelo.
3. Del análisis del estado del arte, se consideró utilizar arquitecturas de transformadores visuales ya que este tipo de redes pueden preservar mejor los detalles finos y evitar artefactos como el *ghosting*, logrando imágenes más nítidas y estables en comparación con las arquitecturas de redes

convolucionales, que tienden a suavizar bordes y perder precisión en el reconocimiento. Por ello, se realizó un entrenamiento mediante aprendizaje por transferencia y se evaluaron más de 10 arquitecturas diferentes de transformadores visuales disponibles en la biblioteca *timm* [4]. Esta evaluación inicial permitió seleccionar la arquitectura con el mejor desempeño, en términos de exactitud.

4. Posterior a la selección de la mejor arquitectura, se procedió a mejorar los parámetros durante el entrenamiento con el fin de optimizar las métricas y lograr una convergencia más estable de la red. Esto incluyó la habilitación de capas *Dropout*, cambios en la función de pérdida (utilizando *Focal Loss* en lugar de *CrossEntropy*) y la modificación del algoritmo optimizador (cambiando de *AdamW* a *AdamP*). También se implementaron épocas de calentamiento con una tasa de aprendizaje baja al inicio del entrenamiento, y se ajustó el programador de tasa de aprendizaje (utilizando *CosineAnnealingLR*). Se exploró el congelamiento de etapas de la red, aunque no se observó una mejora significativa.

En esta evaluación, también se realizaron cambios en el conjunto de entrenamiento, modificando la proporción y el tipo de conjuntos *GAN* incluidos para mejorar la capacidad del modelo de detectar rostros generados por arquitecturas más recientes. Adicionalmente, se aumentó el tamaño de las imágenes de entrada al modelo de 224x224 a 384x384 píxeles para proporcionar una mayor cantidad de información visual y mejorar la extracción de características.

5. Como contribución original, se propusieron y evaluaron dos técnicas de aumento de datos: el aumento de datos basado en el efecto *Deep Dream* y el aumento de datos mediante deformación de imágenes (*warping*). Esta propuesta surgió con base en las características observadas en las imágenes generadas por las GANs y en las incorrectamente mal clasificadas. Estas técnicas de aumento se aplicaron fuera de línea a imágenes del conjunto de rostros naturales para simular artefactos presentes en imágenes generadas y entrenar al modelo para detectarlos. Se ajustó el volumen de los conjuntos de datos de entrenamiento y prueba, incluyendo la generación de imágenes adicionales para los conjuntos *StyleGAN2-ADA* y *StyleGAN3* para asegurar una evaluación más robusta. Como resultado de esta etapa, se generó el modelo final para identificar rostros naturales y sintéticos.
6. Finalmente, el modelo obtenido se evaluó en seis escenarios distintos para determinar su desempeño y robustez en diversas condiciones. Estos escenarios incluyeron:

- a. la evaluación con conjuntos de prueba no vistos durante el entrenamiento (escenario principal).
- b. el análisis de imágenes sometidas a operaciones de post-procesamiento como compresión JPEG, ruido gaussiano, desenfoque y redimensionamiento (escenario 2).
- c. la evaluación con imágenes generadas por modelos GAN que no contienen rostros humanos (escenario 3).
- d. el análisis de imágenes generadas por un modelo GAN 3D (*EG3D*) (escenario 4).
- e. la evaluación con imágenes generadas por un modelo de difusión (*Stable Diffusion*) (escenario 5).
- f. el análisis de imágenes de rostros naturales restaurados mediante un modelo GAN (*GFPGAN*) (escenario 6).

La Figura 1.4 resume la arquitectura final del modelo propuesto. El proceso inicia con una imagen de entrada, la cual se ajusta a un tamaño de 384×384 píxeles para asegurar un balance entre el consumo computacional y la preservación de la información visual.

Posteriormente, las imágenes de entrenamiento pasan por un proceso de aumento de datos que combina dos enfoques. Primero, se aplican técnicas tradicionales en línea, como transformaciones geométricas y de color, para diversificar los datos durante el entrenamiento. Segundo, se utiliza un aumento de datos fuera de línea que constituye una de las contribuciones originales del trabajo: la aplicación de los efectos *Deep Dream* y deformación de imágenes sobre rostros naturales para simular los artefactos sutiles característicos de las imágenes generadas artificialmente.

La imagen procesada se introduce en el modelo predictor, que es la arquitectura MViTv2 con un ajuste de parámetros optimizado. Las mejoras clave en su configuración incluyen la habilitación de una capa de Dropout del 25%, la implementación de épocas de calentamiento para estabilizar la convergencia, el uso del optimizador AdamP, y la aplicación de la función de pérdida Focal Loss para enfocar el aprendizaje en los ejemplos más difíciles. Finalmente, el modelo emite

una clasificación binaria que determina si el rostro en la imagen es natural o fue generado artificialmente.

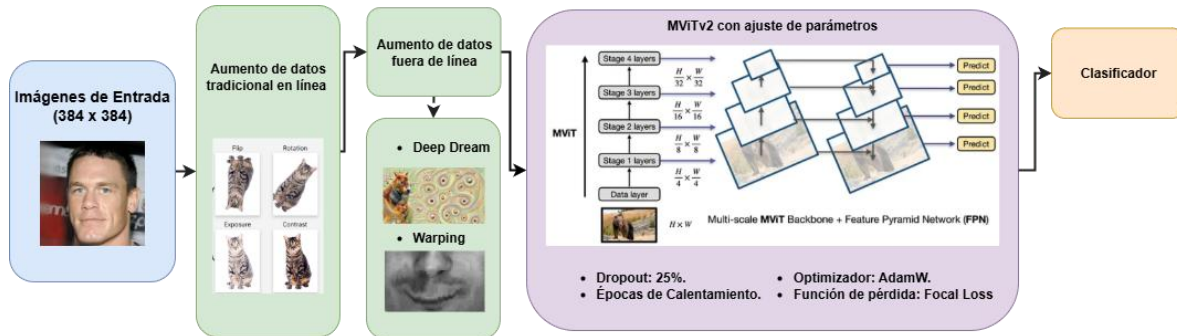


Figura 1.4 Metodología final.

El proceso de inferencia del modelo, ilustrado en la Figura 1.5 detalla el flujo para determinar si una o más imágenes de entrada contienen un rostro natural o uno generado artificialmente. Este proceso comienza con la imagen de entrada, que es la imagen a ser analizada. La imagen se introduce al modelo predictor, que corresponde a la arquitectura *MViTv2* seleccionada y optimizada durante la experimentación. Este modelo, entrenado para la tarea de clasificación binaria, analiza las características de la imagen procesada. Finalmente, el modelo predictor genera una respuesta binaria: "Sí / No" sobre si la imagen fue generada artificialmente. Esta salida indica la predicción del modelo sobre la naturaleza del rostro presente en la imagen de entrada.

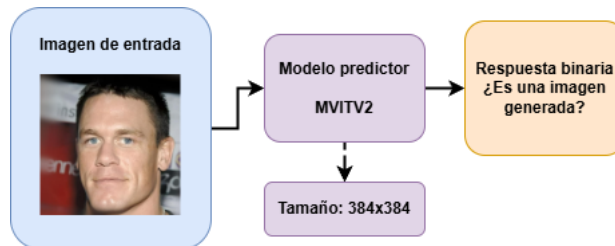


Figura 1.5. Diagrama de inferencia.

1.6. Organización de la tesis

La presente tesis se organiza en siete capítulos, distribuidos de la siguiente manera:

- El Capítulo 1 introduce el problema que motivó esta investigación, estableciendo los objetivos generales y específicos del trabajo, sus alcances y limitaciones, y presentando la metodología propuesta para abordar la problemática.

- El Capítulo 2 se centra en el marco teórico, detallando los fundamentos de visión artificial y aprendizaje profundo, las tecnologías base, las técnicas de generación y manipulación de imágenes, así como las métricas y arquitecturas avanzadas relevantes utilizadas en el desarrollo del proyecto.
- El Capítulo 3 presenta la revisión de la literatura y el estado del arte relacionado con la detección de rostros generados artificialmente, incluyendo trabajos previos desarrollados en el TecNM/CENIDET.
- El Capítulo 4 especifica los conjuntos de datos de rostros naturales y generados artificialmente utilizados en la experimentación.
- El Capítulo 5 contiene la descripción del entorno de hardware y software empleado para el desarrollo y la evaluación del modelo, así como el proceso de experimentación.
- El Capítulo 6 incluye las diferentes pruebas realizadas, los escenarios de evaluación para medir el rendimiento del modelo y el análisis de los resultados obtenidos.
- El Capítulo 7 presenta las conclusiones generales del trabajo, destacando los objetivos logrados, las contribuciones realizadas, las líneas de investigación futuras y las actividades académicas llevadas a cabo durante el desarrollo de la tesis.

Capítulo II: Marco teórico

Este capítulo presenta los fundamentos teóricos y las tecnologías base que sustentan la metodología desarrollada en este trabajo de tesis. Se abordan conceptos clave como visión artificial, el aprendizaje profundo, así como las técnicas relevantes para la generación y manipulación de imágenes. También, se detallan las técnicas de procesamiento y mejora de modelos, las arquitecturas avanzadas empleadas en el desarrollo del modelo propuesto y las métricas utilizadas para la evaluación del desempeño del sistema.

2.1. Fundamentos y tecnologías base

Visión Artificial

La visión artificial es un área de la Inteligencia Artificial que desarrolla algoritmos para reconocer y describir de manera precisa y eficiente a las imágenes. Esta tecnología se basa en el procesamiento digital de imágenes para analizar y extraer información significativa a partir de datos visuales. El objetivo principal de la visión artificial es automatizar tareas visuales altamente repetitivas que sean fatigosas o difíciles de realizar para un operario, especialmente cuando este trabajo es ineficiente o costoso en términos económicos o temporales [5] [6].

El procesamiento digital de imágenes es el conjunto de técnicas y algoritmos que se aplican a imágenes digitales para mejorar su calidad, extraer características de interés, realizar segmentación, detección de objetos, reconocimiento de patrones y mucho más. En la visión artificial, este procesamiento desempeña un papel crucial al permitir que las máquinas procesen y comprendan imágenes de manera eficiente y eficaz.

Las etapas fundamentales para el desarrollo de un sistema de visión artificial se observan en la Figura 2.1, las cuales incluyen [7]:

- **Adquisición de imágenes:** La adquisición puede ser tan sencilla como recibir una imagen ya digitalizada o contar con un dispositivo que realice tal tarea.
- **Procesamiento:** Es el proceso de manipular una imagen para que el resultado sea más adecuado que el original para una aplicación específica. Esta etapa puede incluir escalado, transformaciones en color, etc. En esta etapa también se encuentra la restauración de imágenes ya que es un área que también se

ocupa de mejorar el aspecto de una imagen. Sin embargo, a diferencia de la mejora, que es subjetiva, la restauración de imágenes es objetiva, en el sentido de que las técnicas de restauración suelen basarse en modelos matemáticos o probabilísticos de la degradación de la imagen.

- Segmentación: Divide una imagen en sus partes constituyentes u objetos.
- Representación y descripción: Se ocupa de las herramientas de extracción de componentes de la imagen útiles para la representación y descripción de su contenido ya sea por medio de la forma, color, textura, etc.
- Reconocimiento de objetos: Es el proceso que asigna una etiqueta (por ejemplo, "vehículo") a un objeto a partir de su representación.

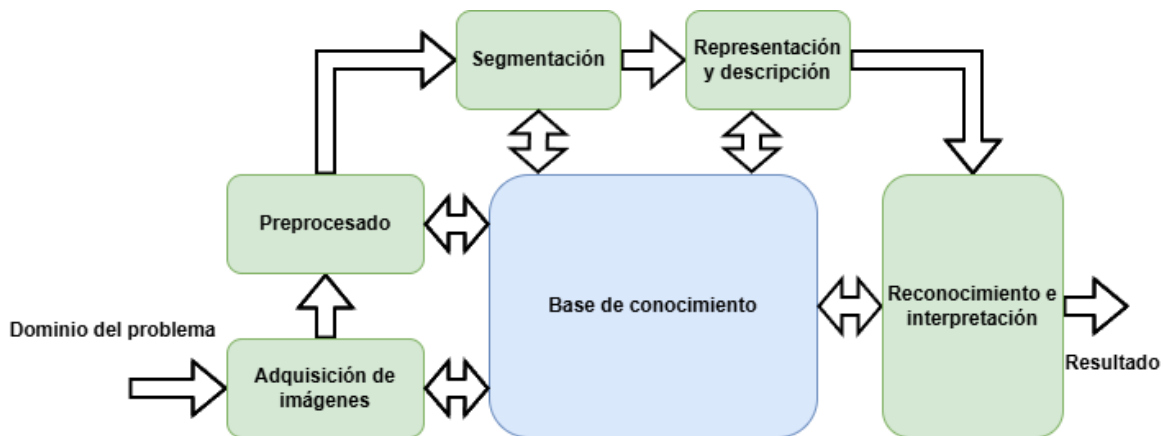


Figura 2.1. Etapas fundamentales del tratamiento digital de imágenes [7].

Deep learning

Los sistemas de reconocimiento de imágenes utilizan redes neuronales artificiales para dotar a las computadoras con la capacidad de “entender las imágenes”. Estos sistemas permiten a las máquinas detectar, identificar y clasificar objetos dentro de las imágenes con bastante precisión. El reconocimiento de imágenes es esencial para diversas áreas comerciales, desde el procesamiento de imágenes digitales simples hasta la interpretación automatizada de dispositivos médicos y militares.

Existen técnicas más modernas basadas en el aprendizaje profundo, que utilizan redes neuronales para aprender a realizar tareas de visión artificial de forma automática a partir de grandes cantidades de datos, las cuales han demostrado ser muy efectivas en tareas como la clasificación de imágenes o la detección de objetos.

El aprendizaje profundo se centra en la generación de representaciones a partir de datos, priorizando la creación de capas sucesivas de representaciones cada vez más significativas. Éstas se aprenden mediante redes neuronales, que consisten en capas literales apiladas una sobre otra.

Los objetivos del aprendizaje profundo son mejorar la capacidad de las máquinas para realizar tareas complejas, como la visión por computadora, el procesamiento del lenguaje natural y la toma de decisiones. Las ventajas del aprendizaje profundo incluyen la capacidad de procesar grandes cantidades de datos, la capacidad de aprender de manera autónoma y la capacidad de mejorar con el tiempo [8].

Existen diferentes arquitecturas del aprendizaje profundo, entre las cuales se encuentran:

- CNN.
- GAN.
- VAE.
- Entre otras.

Algoritmo de optimización

Encargado de ajustar los pesos del modelo en función de la pérdida calculada. Los optimizadores pueden clasificarse en métodos de gradiente estocástico, como *SGD*, y métodos adaptativos como *Adam* y *AdamW*. La elección del optimizador puede influir significativamente en la convergencia del modelo, siendo *AdamW* uno de los más usados en redes profundas debido a su regularización efectiva y manejo del decaimiento de pesos [9].

Dropout

Es una técnica de regularización que desactiva de manera aleatoria neuronas durante el entrenamiento, previniendo el sobreajuste al forzar al modelo a aprender representaciones más robustas y diversas. Este mecanismo es especialmente útil en redes neuronales profundas, donde el riesgo de sobreajuste es mayor debido al alto número de parámetros [9].

Aumento de datos en línea y fuera de línea

El *aumento de datos* es una técnica usada en aprendizaje automático y, especialmente, en redes neuronales para expandir artificialmente el tamaño y la diversidad del conjunto de datos de entrenamiento. Esto es útil cuando se cuenta con datos limitados y se busca mejorar la capacidad de generalización del modelo, evitando el sobreajuste [10].

Aumento de datos en línea, también llamado *on-the-fly* u *online*, aquí, el aumento se realiza durante el proceso de entrenamiento, aplicando transformaciones aleatorias a cada muestra en cada época. Las imágenes o datos se cargan, se aumentan en tiempo real y luego se pasan al modelo. Este método permite un flujo constante de datos variados, reduciendo el espacio de almacenamiento, ya que las

muestras aumentadas no se guardan. Es especialmente útil cuando se entrena con grandes conjuntos de datos o en aplicaciones en tiempo real.

Aumento de datos fuera de línea, también llamado *offline*. En este enfoque, las transformaciones se aplican antes del entrenamiento, generando un conjunto de datos aumentado que se almacena permanentemente. Los datos aumentados se guardan en disco y luego se utilizan para entrenar el modelo. Puede ser conveniente cuando el tiempo de entrenamiento es una prioridad, ya que evita el procesamiento en tiempo real, pero requiere mayor espacio de almacenamiento.

Función de pérdida

Las funciones de pérdida son una parte fundamental en el entrenamiento de modelos de aprendizaje automático, incluyendo las redes neuronales. Estas funciones miden qué tan bien o mal está desempeñándose un modelo en una tarea específica al comparar sus predicciones con los valores reales o esperados. La pérdida obtenida se utiliza como guía para ajustar los parámetros del modelo durante el entrenamiento [9]. *CrossEntropy* y *Focal Loss* son funciones comunes; la primera es adecuada para problemas de clasificación, mientras que la segunda se usa para manejar clases desbalanceadas, enfocándose más en ejemplos difíciles de clasificar.

Épocas de calentamiento

Se refieren a un período inicial en el entrenamiento de un modelo donde se utiliza una tasa de aprendizaje más baja que la habitual [8]. Durante las épocas de calentamiento, la tasa de aprendizaje se incrementa progresivamente hasta alcanzar el valor óptimo o el que se haya configurado para el entrenamiento completo. Esto permite que la red se adapte gradualmente a los datos y reduzca el riesgo de que el modelo quede atrapado en mínimos locales o tenga una alta variabilidad en las primeras iteraciones.

Redes neuronales convolucionales

Las redes neuronales convolucionales son modelos de aprendizaje profundo que se han utilizado ampliamente en aplicaciones de visión artificial. Trabajan aprendiendo una jerarquía de patrones y conceptos modulares para representar el mundo visual.

Las CNN constan de capas apiladas, que realizan operaciones de convolución y agrupación respectivamente. Las capas convolucionales extraen entidades de los datos de entrada mediante la aplicación de filtros para detectar patrones, mientras que las capas agrupadas reducen las dimensiones espaciales de los mapas de entidades y conservan la información más importante [6].

Son conocidas por su capacidad para aprender representaciones complejas a partir de grandes cantidades de datos, lo que las hace adecuadas para tareas como la clasificación de imágenes. Han tenido éxito en el campo de la visión artificial debido a su capacidad para generalizar bien y gestionar con eficacia pequeños conjuntos de datos de entrenamiento.

Biometría facial

La biometría facial es una técnica de reconocimiento que permite identificar o confirmar la identidad de una persona a partir de su rostro. Los sistemas de reconocimiento facial utilizan patrones matemáticos únicos para almacenar los datos biométricos, lo que los hace muy precisos.

El reconocimiento facial se puede utilizar en una variedad de casos de uso, como la verificación de documentos de identidad, el control de acceso a espacios físicos y la banca digital. Además, la biometría facial ofrece más beneficios en comparación con otras técnicas de reconocimiento biométrico, como la biometría de huella dactilar. Por ejemplo, la biometría facial se puede capturar utilizando cámaras de menor resolución disponibles en la mayoría de los teléfonos inteligentes, lo que la hace más accesible y versátil. Además, la biometría facial ofrece resultados precisos sin importar las condiciones de las manos o la presencia de guantes, lo que garantiza una mayor comodidad y eficacia [11].

Es importante tener en cuenta que el uso de la biometría facial también plantea preocupaciones en cuanto a la privacidad y el uso indebido de los datos biométricos.

Algunas de las características que usualmente se analizan en el reconocimiento facial son: la distancia entre los ojos, de la frente a la barbilla, la nariz a la boca, la profundidad de las cuencas oculares, la forma de los pómulos, el contorno de los labios, las orejas y la barbilla, etc.

2.2. Generación y manipulación de imágenes

Red generativa adversaria

Según Goodfellow et al. [12], las redes generativas adversarias, son un marco para estimar modelos generativos a través de un proceso adversarial. Esto implica entrenar dos modelos simultáneamente: un modelo generativo (G) que captura la distribución de datos, y un modelo discriminativo (D) que estima la probabilidad de que una muestra provenga de los datos de entrenamiento en lugar de G .

Los modelos se definen típicamente como perceptrones multicapa, y el procedimiento de entrenamiento involucra un juego mínimo para dos jugadores,

donde G pretende maximizar la probabilidad de que D cometa un error, mientras que D busca clasificar correctamente tanto ejemplos de entrenamiento como muestras de G .

El marco permite la recuperación de la distribución de generación de datos cuando G y D tienen suficiente capacidad, y se puede entrenar usando retropropagación sin la necesidad de cadenas de Markov o redes de inferencia aproximadas desenrolladas.

Deepfakes

Los deepfakes también conocidos como “falsificaciones profundas” son medios, como imágenes o vídeos, que se han manipulado con técnicas de aprendizaje profundo para crear contenido realista pero falso [13].

Por lo general, se generan mediante el uso de GANs y se pueden utilizar para engañar, difundir desinformación, cometer fraudes e interrumpir el funcionamiento del gobierno. Se pueden crear intercambiando los rostros de las personas en los vídeos o alterando el audio para que parezca que alguien está diciendo algo que no ha dicho.

Masood et al. [13] proporciona una descripción detallada de los métodos de falsificaciones profundas de imagen y video existentes, como es vista en la Figura 2.2. A continuación, se realiza una revisión para cada categoría en términos de creación para proporcionar una comprensión más profunda de los diversos enfoques.

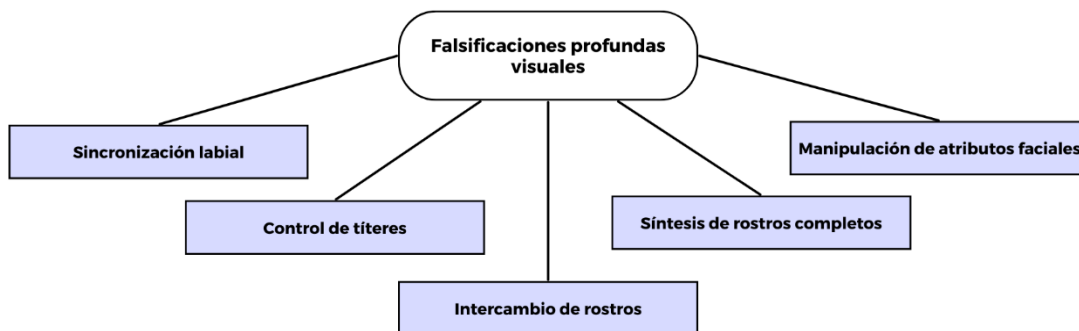


Figura 2.2. Categorización de falsificaciones profundas visuales, adaptada de [13].

Tipos de deepfakes visuales

Se presenta una breve descripción de los tipos de falsificaciones profundas, basada en el trabajo de Masood et al. [13]. La Tabla 2.1 muestra la clasificación de las falsificaciones profundas reportadas y una breve descripción.

Tabla 2.1. Tipos de falsificaciones profundas reportadas, adaptada de [13].

Tipo de falsificación profunda	Descripción
Intercambio de rostros o cambio de identidad	Se trata de un proceso en el que se reemplaza el rostro de una persona por el de otra en un video. El resultado es una manipulación digital convincente, donde la persona original parece ser la persona del rostro sustituido.
Sincronización de labios	Usado para que alguien parezca hablar en un vídeo, aunque en realidad no lo haga. Se sincroniza el movimiento de los labios de la persona en el vídeo con el audio proveniente de otra fuente.
Reactivación de rostros o control de títeres	Consiste en la manipulación del rostro de una persona en un vídeo para que parezca que está realizando expresiones faciales específicas o movimientos predefinidos.
Síntesis de rostros completos	Crea un rostro completamente nuevo, construido desde cero.
Manipulación de atributos faciales	Permite la alteración de ciertos rasgos faciales de una persona en un video, como su edad, género o color de piel.

Para el tema de esta tesis, se abordará la detección de la “síntesis de rostros completos” (cuarto renglón), explorando en profundidad este fenómeno y sus implicaciones.

Síntesis de rostros

La síntesis de rostros completos consiste en crear un rostro completamente nuevo a partir de la nada. En este proceso, se emplean redes generativas de aprendizaje profundo, tales como GAN [12] y VAE [14], para generar imágenes de rostros que se asemejan a la realidad. Para lograr esto, estos modelos son entrenados con amplios conjuntos de datos de imágenes de rostros reales, con el fin de aprender a generar imágenes que sean similares a las de dichos conjuntos de entrenamiento [13]. La Figura 2.3 muestra la mejora en la calidad de generación de rostros artificialmente entre los años 2014 y 2019.



Figura 2.3. Mejora en la calidad de los rostros generados por variaciones de GANs [13].

Es importante destacar que, aunque la generación de rostros mediante IA ha avanzado significativamente, existen ciertos fallos en el proceso que, aunque rara vez ocurren, siguen siendo un desafío. Sin embargo, con cada nueva versión de las arquitecturas y avances en la tecnología, estos fallos van disminuyendo progresivamente. En la Figura 2.4 se exploran ejemplos de estas rarezas o fallos en la generación que denotan que la imagen es creada artificialmente.

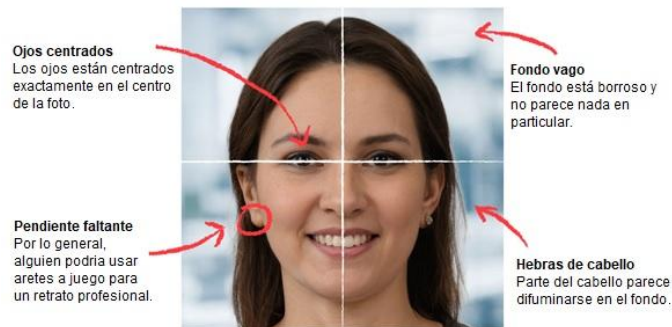


Figura 2.4. Rarezas en la imagen sintetizada [1].

La síntesis de rostros completos ha encontrado aplicaciones en diversas áreas, tales como la creación automática de personajes para videojuegos y en la industria del modelado de rostros en 3D. No obstante, también se ha utilizado con propósitos maliciosos, como la creación de perfiles falsos en redes sociales para la propagación de desinformación [1], [15].

A pesar de la capacidad de la síntesis de rostros completos para generar imágenes realistas, aún se enfrenta a desafíos para lograr una representación convincente. Por ejemplo, la generación de imágenes de rostros que luzcan realistas desde diferentes ángulos o con diversas expresiones faciales puede resultar complicada [16], [17].

Una de las técnicas más comunes para detectar la síntesis de rostros implica la identificación de artefactos. Frecuentemente, las imágenes de rostros sintetizados

presentan defectos visuales, como bordes borrosos o distorsiones en la imagen, que pueden ser identificados mediante algoritmos de detección de artefactos [18], [19].

Modelos de difusión

Son una clase de modelos generativos que funcionan mediante la adición progresiva de ruido gaussiano a los datos originales y el entrenamiento de una cadena de Markov que pueda revertir este proceso, un ejemplo se puede ver en la Figura 2.20. Este tipo de modelo está diseñado para degradar gradualmente la información de la muestra en pasos sucesivos, permitiendo que cada transición añada ruido a los datos originales. Al final de este proceso, la muestra está completamente transformada en ruido aleatorio, esencialmente eliminando cualquier rastro del dato inicial. Sin embargo, el modelo está entrenado para deshacer este proceso: a partir de una imagen de ruido puro, el modelo aplica sucesivos pasos en la dirección opuesta para "revertir" la degradación y reconstruir una muestra que se asemeja a los datos originales [20].

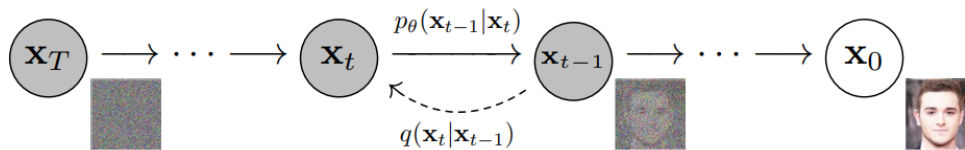


Figura 2.20. Representación gráfica de la metodología [26].

El proceso de entrenamiento de un modelo de difusión implica definir dos componentes: un proceso de avance (o de difusión) y un proceso de reversión. El proceso de avance está predefinido y es el encargado de añadir ruido a la muestra en cada paso, mientras que el proceso de reversión es el que el modelo aprende durante el entrenamiento. Para que esto funcione, el modelo se entrena con inferencia variacional, optimizando una serie de distribuciones gaussianas en cada paso que permitan reconstruir la muestra. De este modo, el modelo aprende a generar muestras nuevas de manera muy similar a los datos originales.

Durante la generación, el modelo inicia con una imagen completamente ruidosa y, paso a paso, aplica su proceso aprendido para ir eliminando el ruido en cada etapa hasta llegar a una muestra coherente y estructurada. Esta estructura de ruido progresivo permite que los modelos de difusión realicen una compresión de información de tipo progresivo y cuenten con características específicas para trabajar en tareas generativas, como la síntesis de imágenes.

Stable Diffusion [21]

Es una metodología para la síntesis de imágenes de alta resolución basada en modelos de difusión latentes. Estos modelos descomponen el proceso de generación de imágenes en una serie de pasos secuenciales que emplean autoencoders de eliminación de ruido. Esta estrategia permite controlar el proceso de generación de imágenes sin necesidad de entrenar nuevamente el modelo.

Una de las características clave es que no opera directamente en el espacio de píxeles, como lo hacen otros modelos de difusión. En su lugar, trabaja en un espacio latente más eficiente generado por autoencoders previamente entrenados. Este enfoque reduce significativamente los requisitos computacionales y, al mismo tiempo, mejora la calidad y fidelidad visual de las imágenes generadas.

Se destacan por su capacidad para realizar tareas como completar imágenes, super-resolución, síntesis de imágenes condicionales y generación de imágenes de texto a imagen, con un menor costo computacional en comparación con métodos anteriores. Esto se logra al combinar la compresión de datos perceptuales con modelos generativos robustos en el espacio latente, optimizando la relación entre la complejidad y la preservación de detalles visuales.

Deep Dream

Es un proyecto de Google, basado en redes neuronales profundas, que transforma imágenes mediante un proceso que resalta patrones y características visuales. Fue desarrollado inicialmente para entender mejor cómo las redes neuronales, específicamente las redes convolucionales, reconocen y procesan características en imágenes. El proceso detrás de *Deep Dream* involucra la *alucinación* de patrones y formas que una red ya entrenada identifica dentro de una imagen, amplificando lo que "cree" ver. Esto genera efectos visuales abstractos o surrealistas, donde texturas, colores y detalles se realzan y multiplican [22].

El funcionamiento de *Deep Dream* es el siguiente:

1. Se utiliza una red neuronal convolucional previamente entrenada en una tarea de clasificación de imágenes, como la red *Inception* de Google, que está diseñada para identificar objetos comunes.
2. En lugar de realizar una predicción normal, el algoritmo utiliza la retroalimentación para *amplificar* ciertas características o patrones que detecta en una imagen. Se elige una capa de la red y, mediante retropropagación, se ajustan los valores de los píxeles para maximizar las activaciones en esa capa.
3. A medida que la red intenta realzar ciertos patrones que "ve" en la imagen, se crean repeticiones, texturas y formas inusuales que producen el aspecto

característico de una imagen *Deep Dream*, con detalles tipo espirales, ojos o motivos animales.

Las imágenes generadas por *Deep Dream* suelen tener un aspecto psicodélico y surrealista, con patrones repetitivos que emergen de la estructura original de la imagen. Los objetos pueden aparecer distorsionados, con formas que recuerdan a ojos, animales o paisajes alienígenas, lo que les da un aspecto onírico.

Deformación de imágenes

La deformación de imágenes, también conocida como *image warping*, es una técnica de procesamiento de imágenes que modifica la forma o estructura de una imagen para ajustarla a una nueva geometría. Este proceso se utiliza en diversas aplicaciones, desde la corrección de distorsiones ópticas y la creación de efectos visuales hasta la generación de mapas de calor o el análisis de movimiento en visión por computadora [23].

El proceso implica transformar la posición de los píxeles en la imagen original para que se alineen con una nueva disposición. La técnica aplica una función de transformación matemática que modifica las coordenadas de los píxeles de acuerdo con los requisitos del efecto deseado. Existen varias transformaciones comunes:

- Transformación afín: Mantiene paralelas las líneas rectas y permite realizar operaciones como rotación, escalado, traslación y sesgado. Es útil para ajustar la orientación y el tamaño de una imagen.
- Transformación de perspectiva: Permite crear un efecto de profundidad, dando la impresión de que la imagen se está viendo desde un ángulo diferente. Esto es común en aplicaciones de corrección de perspectiva, como la corrección de fotografías tomadas en ángulo.
- Deformación no lineal: En esta transformación, los píxeles se mueven en función de funciones matemáticas más complejas, permitiendo curvas y distorsiones de mayor flexibilidad. Esto se usa para efectos visuales, como hacer que una imagen parezca "líquida" o para ajustar una imagen a una forma no rectangular.
- Interpolación: Cuando se realiza una deformación, se deben calcular los valores de los nuevos píxeles en las posiciones modificadas. La interpolación se encarga de estimar los valores de estos píxeles basándose en los píxeles vecinos, manteniendo una imagen suave y continua.

En el contexto de este trabajo, es crucial entender estas transformaciones comunes para explicar cómo la deformación de imágenes se aplica como técnica de aumento de datos. Al someter las imágenes a estos cambios geométricos, se incrementa la diversidad del conjunto de datos de entrenamiento, lo cual permite

al modelo aprender a reconocer patrones de rostros generados artificialmente de manera más robusta, independientemente de la orientación, la perspectiva o las distorsiones, mejorando así su capacidad de generalización y su desempeño.

2.3. Métricas de rendimiento

Matriz de confusión

Un método común para describir el rendimiento de un modelo de clasificación es la matriz de confusión [24].

La matriz de confusión es una simple tabulación cruzada entre lo observado y las clases previstas para los datos. La Tabla 2.2 muestra un ejemplo cuando el resultado tiene dos clases. Las celdas diagonales denotan casos en los que las clases son predichas correctamente, mientras que las celdas que no están en la diagonal representan los errores de predicción para cada caso posible.

Tabla 2.2. Matriz de confusión para dos clases. Las celdas indican el número de positivos verdaderos (TP), falsos positivos (FP), negativos verdaderos (TN) y falsos negativos (FN) [24].

Predicción	Observación	
	Positivo	Negativo
Positivo	TP	FP
Negativo	FN	TN

En donde:

- Verdadero positivo (TP): Son los casos en los que los datos reales y la predicción son positivos.
- Verdadero negativo (TN): Son los casos en los que los datos reales y la predicción son negativos.
- Falso positivo (FP): Son los casos en los que la predicción es positiva pero los datos reales son negativos.
- Falso negativo (FN): Son los casos en los que la predicción es negativa pero los datos reales son positivos.

A partir de la matriz de confusión, se derivan diversas métricas de evaluación [25], que incluyen la exactitud (ecuación 1), la sensibilidad (ecuación 2), la especificidad (ecuación 3), la precisión (ecuación 4) y la puntuación F1 (ecuación 5). Estas métricas se presentan a continuación.

Exactitud

La exactitud es el porcentaje general de datos clasificados de manera correcta. Se calcula dividiendo el número de predicciones correctas, tanto positivas como negativas, entre el total de predicciones.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Sensibilidad

Es comúnmente denominada la tasa de verdaderos positivos o recuperación y representa el porcentaje de casos positivos que el algoritmo identificó de manera precisa. Esto demuestra la habilidad del algoritmo para distinguir entre casos positivos y negativos.

$$TPR = \frac{TP}{TP + FN} \quad (2)$$

Especificidad

También llamada tasa de verdaderos negativos, es el porcentaje de casos negativos que el algoritmo identifica de manera precisa. Refleja la capacidad del modelo para distinguir casos negativos de positivos.

$$TNR = \frac{TN}{FP + TN} \quad (3)$$

Precisión

Porcentaje de clasificación correcta cuando se predice positivo. Se refiere a lo cerca que está el resultado de una medición del valor verdadero. Se representa por la proporción entre los positivos reales predichos por el algoritmo y todos los casos positivos.

$$P = \frac{TP}{TP + FP} \quad (4)$$

Puntuación F1

La puntuación F1 [26] es una medida de la precisión de un modelo que tiene en cuenta tanto la precisión como la recuperación. Es la media armónica de precisión y recuperación, que da igual peso a ambas métricas. La puntuación F1 proporciona un valor único que resume el rendimiento del modelo en términos de precisión y

recuperación. La puntuación de la F1 oscila entre 0 y 1, y una puntuación más alta indica un mejor rendimiento.

$$F1 = 2 \left(\frac{PR * TPR}{PR + TPR} \right) \quad (5)$$

Curva ROC [27]

La curva característica operativa del receptor constituye una herramienta importante para evaluar el rendimiento de un modelo de *machine learning*. Por lo general, se utilizan en problemas de clasificación binaria, concretamente, problemas con dos clases de salidas distintas. La curva ROC muestra la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos del modelo. *TPR* describe la tasa a la que el clasificador predice como “positivo” a las observaciones que son “positivas”. *FPR* describe la tasa a la que el clasificador predice como “positivo” a las observaciones que en realidad son “negativas”. Los clasificadores perfectos tienen una *TPR* de 1 y una *FPR* de 0.

2.4. Modelos de transformadores visuales

Transformador visual [28]

Arquitectura diseñada para abordar el reconocimiento de imágenes utilizando el enfoque de los Transformers, eliminando la necesidad de las tradicionales redes convolucionales. En lugar de aplicar convoluciones para extraer características de las imágenes, este modelo utiliza el mecanismo de atención sobre bloques de la imagen.

Consiste en dividir una imagen en pequeños parches, generalmente de tamaño 16x16 píxeles. Cada uno de estos parches se trata como una "palabra" en el sentido de que se convierte en una representación lineal mediante una proyección. Estos vectores resultantes se procesan como una secuencia, de forma similar a cómo se manejan las palabras en tareas de procesamiento de lenguaje natural. Así, el modelo puede aplicar atención auto-regresiva o "self-attention" entre los diferentes parches de la imagen para capturar relaciones espaciales y de contenido.

La arquitectura (Figura 2.5) consiste en varias capas de atención múltiple, donde se implementan múltiples cabezas de atención en paralelo, permitiendo al modelo aprender diferentes representaciones de los parches simultáneamente. Cada capa incluye también mecanismos de normalización y conexiones residuales, lo que ayuda a estabilizar el entrenamiento y permite que el modelo maneje con eficiencia

las imágenes de alta resolución. Después de estas capas de atención, la salida se pasa por un clasificador lineal que asigna la imagen a una categoría específica. [28]

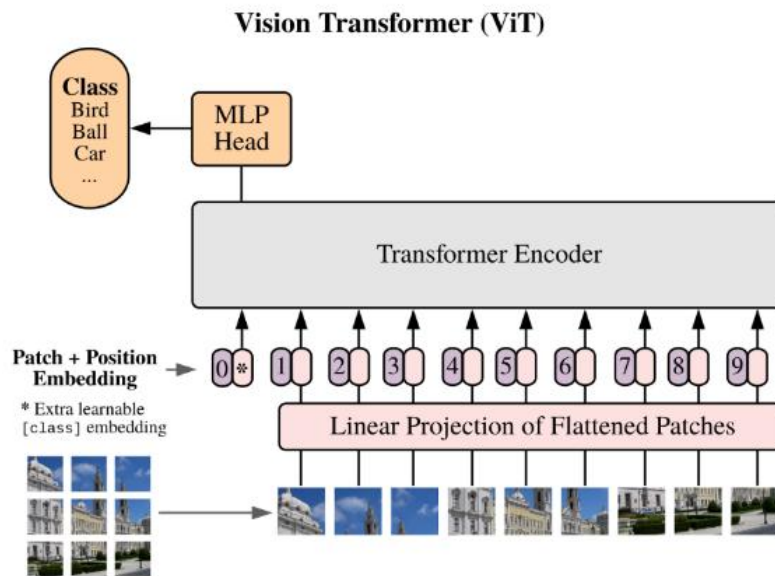


Figura 2.5 Vista general del modelo ViT [28].

Dado que las imágenes no contienen una estructura secuencial implícita como los datos textuales, el modelo incorpora codificaciones posicionales para agregar información sobre la disposición espacial de los parches. Estas codificaciones permiten que el modelo reconozca la relación entre los parches en función de su ubicación relativa, lo cual es crucial para una interpretación efectiva de las imágenes.

MViTv2 (Improved Multiscale Vision Transformers for Classification and Detection) [29]

La arquitectura *MViTv2* es una versión mejorada de los transformadores de visión multiescala [30], diseñada para tareas de clasificación de imágenes, detección de objetos y reconocimiento de videos.

La arquitectura de *MViTv2* para detección de objetos utiliza un extractor de características multiescala, la cual se observa en la Figura 2.6, que se integra con una red piramidal de características. La estructura está organizada en cuatro etapas (*stages*), cada una de las cuales tiene un nivel de resolución diferente, lo que permite extraer características de múltiples escalas de la imagen de entrada.

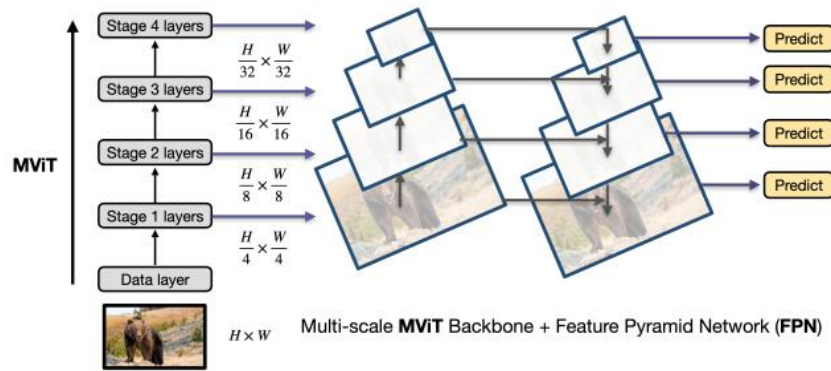


Figura 2.6 Red piramidal de características [29].

En la etapa 1, se procesa la imagen de entrada, reduciendo la resolución a un cuarto ($1/4$) de la original. Esta etapa realiza operaciones de agrupamiento inicial, lo que permite reducir la dimensión espacial mientras se mantienen características clave. Esto permite al modelo captar información fundamental sobre la imagen en su nivel más básico, estableciendo una base sólida para el procesamiento posterior.

La etapa 2 utiliza la salida de la etapa anterior como entrada, reduciendo aún más la resolución, concretamente a un octavo ($1/8$) de la original. En esta fase, el modelo sigue capturando características de nivel medio, pero de manera más compacta. Aquí, se comienza a representar de manera más eficiente las estructuras y patrones que aparecen en la imagen, sin perder la conexión con el contexto espacial general.

En la etapa 3, la resolución se reduce a un dieciseisavo ($1/16$) del tamaño original. Esta etapa se enfoca en modelar características más abstractas y de mayor nivel, lo cual es esencial para la identificación de estructuras y patrones más complejos en la imagen. El procesamiento en esta fase permite una comprensión más profunda de las relaciones espaciales y de las características presentes en la imagen.

La etapa 4 reduce la resolución a un treintaidosavo ($1/32$) de su tamaño original y se centra en características de nivel más alto. Para el reconocimiento de objetos a gran escala o en posiciones específicas dentro de la imagen, dado que las características semánticas que se generan son más detalladas y adecuadas para tareas de detección precisas.

La integración de *MViTv2* con la *FPN* facilita la combinación efectiva de las características multiescala generadas en cada etapa. La *FPN* utiliza una estructura de conexión piramidal de arriba hacia abajo, lo que permite fusionar características de diferentes niveles de resolución. Esto genera mapas de características semánticamente coherentes para cada escala, los cuales son utilizados para predecir la presencia y ubicación de objetos en la imagen.

La arquitectura se implementa en cinco tamaños diferentes: *Tiny (T)*, *Small (S)*, *Base (B)*, *Large (L)* y *Huge (H)*. Cada uno de estos modelos se adapta a diferentes necesidades en términos de recursos computacionales y rendimiento.

El modelo *T* es la versión más liviana, diseñada para ser más rápida y eficiente en términos de consumo de memoria. Es adecuada para aplicaciones en las que la capacidad computacional es limitada y se necesita una respuesta rápida, aunque con menor precisión en comparación con los modelos más grandes.

Los modelos *B*, *L* y *H* se enfocan en lograr altos niveles de precisión y están diseñados para tareas de visión más complejas, como detección de objetos de alta resolución y reconocimiento de videos. Estos modelos tienen más bloques, canales y cabezales de atención, como se ve en la Figura 2.7, lo que les permite capturar detalles más finos y características semánticas de alto nivel. Sin embargo, estos modelos son más demandantes en términos de recursos computacionales y memoria.

Model	#Channels	#Blocks	#Heads	FLOPs	Param
MViT-T	[96-192-384-768]	[1-2-5-2]	[1-2-4-8]	4.7	24
MViT-S	[96-192-384-768]	[1-2-11-2]	[1-2-4-8]	7.0	35
MViT-B	[96-192-384-768]	[2-3-16-3]	[1-2-4-8]	10.2	52
MViT-L	[144-288-576-1152]	[2-6-36-4]	[2-4-8-16]	39.6	218
MViT-H	[192-384-768-1536]	[4-8-60-8]	[3-6-12-24]	120.6	667

Figura 2.7 Configuración de las variantes de MViTv2 [29].

Para mejorar el rendimiento en tareas de clasificación y detección, se aplica ajuste fino a los modelos *S*, *B* y *L* (Figura 2.8) utilizando conjuntos de datos de ImageNet, como *ImageNet-1K*, habiendo utilizado *ImageNet-21K* para su entrenamiento previo. El ajuste fino con *ImageNet* permite que estos modelos preentrenados aprovechen el conocimiento adquirido sobre una amplia gama de imágenes, mejorando su capacidad de generalización para tareas más específicas. Por ejemplo, el ajuste fino de *MViTv2-B* con *ImageNet-1K* aumenta su precisión en clasificación de imágenes. También se reportan pruebas de la misma índole con el modelo *H*, sin embargo, solo se realizan pruebas con la detección de objetos.

model	$T \times \tau$	scratch	IN1k	IN21k	FLOPs	Param
MViTv2-S	16×4	81.2	82.2	82.6	64	34.5
MViTv2-B	32×3	82.9	83.3	84.3	225	51.2
MViTv2-L	40×3	81.4	83.4	84.5	1127	217.6
MViTv2-L $\uparrow$ 312 ²	40×3	81.8	84.4	85.7	2828	217.6

Figura 2.8 Efectos del preentrenamiento con ImageNet en los modelos [29].

Capítulo III:

Estado del arte

El capítulo incluye la revisión de la literatura en relación con las técnicas de diferenciación entre rostros naturales y generados artificialmente, principalmente por modelos GAN. También se incluyen trabajos previos desarrollados en el TecNM/CENIDET que están relacionados con la identificación de rostros.

3.1. Trabajos relacionados

En Departamento de Ciencias Computacionales del CENIDET cuenta con varias tesis que utilizan como base el rostro, entre ellas se encuentran las siguientes:

- “*Aplicación de celdas holográficas sobre mapas de detalle para el reconocimiento de rostros*” presentado por Larissa Jeanete Peniche Ruiz [31].
- “*Metodología para el Reconocimiento de Fonemas en el Idioma Español mediante Percepción Visual por Computadora*” presentado por María Guadalupe Medina Barrera [32].
- “*Reconocimiento de rostros invariante a expresiones faciales*” presentado por Andrea Magadán Salazar [33].
- “*Seguimiento de objetos alfaflexibles en secuencias de imágenes*” presentado por Anely Herrera Peralta [34].
- “*Seguimiento y Caracterización de Componentes del Rostro para la Detección de Expresiones Faciales*” presentado por Daniel Alonso Paredes Soto [35].
- “*Método para el reconocimiento de rostros mediante parámetros fractales y RNA’s*” presentado por Miguel Ángel Zúñiga García [36].
- “*Caracterización de Expresiones Faciales con un Sensor RGBD*” presentado por Rita Alejandra Morfin Villatoro [37].
- “*Valoración automática de la motivación y atención del estudiante mediante expresiones faciales*” presentado por Catalina Alejandra Vázquez Rodríguez [38].
- “*Reconocimiento automático del rostro para verificación de identidad para evaluación en línea*” presentado por Wendy Valderrama Cárdenas [39].
- “*Detección de suplantación de identidad en imágenes faciales*” presentado por Wendy Valderrama Cárdenas [40].

- “*Mejoramiento de imágenes para sistemas de videovigilancia de baja calidad*” presentado por Rebecca Yuziel Zumaya Lanuza [41].
- “*Sistema de Visión Artificial para el Reconocimiento de la Mascarilla Facial Bien Puesta*” presentado por Omar de Jesus Trujillo Quintero [42].
- “*Identificación de patrones atípicos en personas con movimientos involuntarios*” presentado por Darien Calcedo Aguirre [43].

Las tesis aportan conocimiento sobre qué características pueden utilizarse para describir el rostro, la efectividad de las mismas, las técnicas recomendadas para la localización de la cara y sus componentes, etc. Sin embargo, cabe resaltar que en todas las tesis anteriores las imágenes utilizadas fueron adquiridas de personas reales. Y el presente trabajo aborda el problema de cómo describir un rostro generado artificialmente que permita una diferencia con los rostros naturales.

3.2. Estado del arte

Se recopiló información sobre 44 artículos relacionados al tema de detección de rostros generados artificialmente mediante el enfoque del análisis de visión artificial. Los artículos se encuentran categorizados en las siguientes tablas:

- Tabla 3.1: presenta los trabajos sobre detección de imágenes y rostros generados artificialmente, que detallan los avances técnicos, las metodologías y los resultados de experimentación
- Tabla 3.2: integra estudios de revisión sobre la detección de imágenes y rostros generados artificialmente, debido a que permiten entender diversas perspectivas, proporcionando datos cuantitativos que muestran tendencias y desafíos comunes.
- Tabla 3.3: resume algunas de las principales aplicaciones y bibliotecas utilizadas para obtener imágenes de rostros generados artificialmente.

Posterior a la síntesis tabular de los artículos revisados, se proporciona una recapitulación de las principales contribuciones y avances dentro del estado del arte.

Tabla 3.1. Síntesis tabular de los trabajos examinados relacionados al reconocimiento de rostros generados.

Referencia	Técnicas	Conjuntos de datos	Métricas	Resultados
2019 [44]	KNN	CelebA, ProGAN, Glow	AUROC	AUROC: 85.2% (más alto)
2019 [45]	CNN*	CelebA, LSUN, ProGAN, SNGAN, CramerGAN, MMDGAN	ACC	ACC: 99.93% (en el conjunto CelebA).
2019 [46]	CNN*	CelebA-HQ, PGGAN	ACC	ACC: 97.97%
2019 [47]	Clasificador Incremental CNN	Cycle-GAN, ProGAN, Glow, starGAN	ACC	ACC: 98.28% (más alto)
2020 [48]	CNN*	PGGAN, StyleGAN2, starGAN, STGAN, FF++,DFDC,Celeb-DF	P, TPR, F1, ACC, AP (precisión promedio), AUC, FPR, FNR	P: 98.6%, TPR: 92.6%, F1: 98.6%, ACC: 98.6%, AP: 97.9%, AUC: 98.5%, FPR: 0.013, FNR: 0.016
2020 [49]	CNN*	PGGAN, TLGAN	ACC	ACC: 80% (aproximado)
2020 [50]	CNN (Gram-Net)	StyleGAN, CelebA-HQ, PGGAN, FFHQ	ACC	ACC: 98.96% (más alto)
2020 [51]	CNN (ResNet-50)	ProGAN, StyleGAN, BigGAN, CycleGAN, StarGAN, GauGAN, CRN, IMLE, SITD, SAN, FF++	AP	AP: 99.6% (más alto)
2020 [52]	CNN (Xception y ForensicsTransfer)	CelebA-HQ, FFHQ, starGAN, GLOW, ProGAN, StyleGAN	ACC	ACC: 99.9% (más alto)
2020 [53]	CNN* (Clasificador por parches)	PGGAN, StyleGAN, StyleGAN2, GLOW, GMM	AP	AP: 97.32%
2020 [54]	CNN* (T-GD)	PGGAN, starGAN, StyleGAN, StyleGAN2	AUROC	AUROC: 98.39% (más alto)
2020 [55]	CNN*	FFHQ, StyleGAN, StyleGAN2	ACC	ACC: 99.70%
2021 [56]	CNN*	StyleGAN2, CycleGAN	ACC	ACC: 98.4%
2021 [57]	CNN*	CelebA-HQ, FFHQ, FF+, C-DF, PGGAN, StyleGAN, StyleGAN2, starGAN, starGANv2	ACC, AU ¹ ROC	ACC: 98.65%, AUROC: 99.90%
2022 [58]	CNN*	PGGAN, StyleGAN, StyleGAN2, StyleGAN2-ADA	ACC	ACC: 89.38%
2022 [59]	CNN	StyleGAN2, VIPPrint,FFHQ	ACC	ACC: 99.53% (más alto)
2022 [60]	CNN*	PGGAN, StyleGAN, StyleGAN2, starGAN, CycleGAN, BigGAN	ACC	ACC: 99.8% (más alto)
2023 [61]	Deep Forest	StyleGAN, StyleGAN2, StyleGAN3, FFHQ	ACC	ACC: 100% y 98.7% (más altos)

¹ * significa que es diseñada en el trabajo.

Tabla 3.1. Síntesis tabular de los trabajos examinados relacionados al reconocimiento de rostros generados (continuación).

Referencia	Técnicas	Conjuntos de datos	Métricas	Resultados
2019 [62]	SVM	LSUN, ImageNet, CelebA	AUC	AUC: 92%
2019 [63]	SVM	ProGAN / PGGAN, CelebA	AUC	AUC: 94%
2019 [64]	CNN	CycleGAN, ProGAN / PGGAN	AUC	AUC: 99%.
2019 [65]	CNN	CycleGAN, StarGAN, CelebA	ACC	ACC: 97.84%
2021 [66]	CNN	CelebA, CelebA-HQ, FaceForensics, ProGAN / PGGAN, StyleGAN, Face2Face, Glow, StarGAN	ACC	ACC: 98.52%.
2021 [67]	Mejora de Xception	Propio, basado en reconstrucción de FFHQ	ACC, P, TPR	ACC: 99.25%, P: 99.31%, TPR: 99.16%
2021 [68]	Residual Attention Network	Conjunto propio (FFHQ-GAN), FFHQ, StyleGAN2	ACC, P, TPR, F1, AUC	Valores más altos: ACC: 97%, P: 98%, TPR: 97%, F1: 97%, AUC: 100%.
2021 [69]	Local Binary Pattern - Net	FFHQ, StyleGAN	ACC, P, TPR	A: 98.58%, P: 98.96%, R: 98.21%
2021 [70]	Modificación de CNN Xception	ProGAN / PGGAN, StarGAN, CycleGAN, StyleGAN, SPADE/GauGAN	ACC	ACC: 98.17%
2022 [71]	Mejora de Xception	CelebA, ProGAN / PGGAN	ACC	ACC: 98.5%.
2022 [17]	SSN	CelebA-HQ, FFHQ, StyleGAN2, StyleGAN3, ProGAN / PGGAN	AUC, TPR/TNR	AUC: 100% en condiciones coincidentes y 99.6% en condiciones no coincidentes.
2022 [72]	Multi-Channel Convolutional Neural Network (MCCNN)	FFHQ, StyleGAN, ProGAN / PGGAN, StyleGAN2, BigGAN, Deepfake, DCGAN, VQ-VAE 2.0	ACC, PR, TPR, F1	ACC: 93.1%, PR: 93.5%, TPR: 99.4%, F1: 96.4%.
2022 [73]	Two Stream CNN	StyleGAN, StyleGAN2, ProGAN / PGGAN, WGAN, BeGAN, LsGAN	ACC	ACC: 99.04%
2022 [74]	CNN	Recopilado / Propio	ACC de validación	ACC: 95.21%
2022 [75]	Combinación ResNet50 y red propia	FFHQ, StyleGAN, ProGAN / PGGAN, StyleGAN2, BigGAN, Deepfake, DCGAN, VQ-VAE 2.0, entre otros	P	P: 96.906%
2022 [18]	BloU	FFHQ, StyleGAN2	AUC	AUC: 91%
2023 [76]	CNN*	CelebA-HQ, PGGAN, CramerGAN, SNGAN, MMDGAN, StyleGAN, StyleGAN2	ACC	ACC: 100% y 98.9% (más altos)
2023 [77]	CNN (Xception)	CelebA-HQ, StyleGAN,	ACC	ACC: 98.77%
2023 [78]	CNN (Xception)	PGGAN, StyleGAN, StyleGAN2, BigGAN, CycleGAN, starGAN, GauGAN	ACC	ACC: 99.4%
2023 [79]	Combinación ResNeSt-14, EfficientNet-B5, U-net y red propia	ProGAN / PGGAN, FFHQ, StyleGAN2	AUC, ACC	AUC: 99.8% (más alto entre pruebas realizadas)
2023 [16]	Inversión GAN	FFHQ, CelebA, CelebHQ, LFW, Caltech, StyleGAN2	AUC	AUC: 88%
2023 [80]	Modificación de Xception	CelebA, LFW	AUC	AUC: 93.2%

Tabla 3.2. Síntesis tabular de los artículos de revisión examinados.

Referencia	Número de trabajos analizados ²	Periodo	Comentarios
2020 [81]	12	2018 - 2020	No contiene evaluaciones en conjuntos de datos posteriores a ProGAN.
2021 [82]	20	2018 - 2020	Realiza evaluación de 7 trabajos sobre detección de imágenes generadas artificialmente.
2022 [83]	40	2011 - 2021	Incluye investigación sobre detección de Deepfakes y conjuntos de datos.
2023 [84]	27	2018 - 2023	Proporciona breve cronología sobre redes generadoras y trabajos en detección de Deepfakes.

Tabla 3.3. Herramientas populares generadoras de rostros.

Nombre	Arquitectura generadora	Código fuente	Tipo	Accesibilidad
DragGAN [85], [86]	DragGAN	Público	Aplicación	Gratuito
Face generator [87]	Privado	Privado	Aplicación	Pago / Prueba gratuita
Anonymizer [87]	Privado	Privado	Aplicación	Pago / Prueba gratuita
Playform FaceMix [88]	Privado	Privado	Aplicación	Pago / Prueba gratuita
Face Studio [89]	Privado	Privado	Aplicación	Pago / Prueba gratuita
Fotor [90]	Privado	Privado	Aplicación	Pago / Prueba gratuita
Thispersondoesnotexist [91]	StyleGAN2	Público	Aplicación	Gratuito

² Se contabilizan los trabajos sobre detección de rostros generados artificialmente, así como otros tipos de *Deepfakes*.

Análisis del estado del arte

Se puede clasificar la detección de rostros falsos en tres enfoques principales: aquellos basados en el aprendizaje profundo, las características físicas y las características fisiológicas.

Dentro del enfoque de características físicas, se destacan los trabajos de Hu et al. [19] y Shinde et al. [92], los cuales examinan la incoherencia de los reflejos especulares corneales entre ambos ojos. No obstante, estos métodos operan bajo supuestos rigurosos, como la presencia de caras en posición frontal o la visibilidad clara de un reflector en los ojos.

En el enfoque fisiológico, las investigaciones se centran en aspectos semánticos de los rostros humanos, abordando elementos como la simetría [17], [63], el color del iris [62], formas de las pupilas [18], [79], entre otros. En cuanto a los enfoques basados en características, se encuentran trabajos como el de Guo et al. [18] donde se focaliza en la forma de la pupila. En este contexto, se espera que la pupila sea, generalmente, circular en la superficie ocular o aparezca de manera elíptica en la imagen en casos de observación con orientaciones distintas a la frontal.

Cabe mencionar que las metodologías que dependen del análisis de atributos físicos tales como [17], [18], [19], enfrentan restricciones derivadas de condiciones ambientales variables, como perspectivas diversas, condiciones de iluminación cambiantes y variaciones en la posición de los sujetos. Además, la presencia de oclusiones en las áreas sujetas a análisis, como gafas, vello facial o accesorios, constituye un desafío adicional para dichos enfoques.

En trabajos como el de Hu et al. [19], encuentran que los rostros generados por la primera arquitectura *StyleGAN* [3] tienen considerables artefactos como huellas en la generación [64], [73], colores de iris inconsistentes [62], etc. Sin embargo, sólo un año después, *StyleGAN2* es propuesto por Karras et al. en [2], mejorando enormemente la calidad visual y la resolución de píxeles en los rostros generados. Un año después de esta arquitectura, se presenta *StyleGAN3* [93], la cual tiene una mejor compatibilidad con conjuntos de datos no alineados y mejora la naturalidad de las imágenes generadas. Es relevante señalar que, incluso en los trabajos más recientes, el conjunto de datos generado por *StyleGAN3* no se utiliza de manera generalizada, con la excepción del trabajo realizado por Wang et al. [17].

Yang et al. [63] descubrió que las GANs tienen la capacidad de generar partes faciales, como ojos, nariz, piel y boca, con un nivel detallado de realismo. Sin embargo, no existe una restricción explícita sobre las ubicaciones específicas de estas partes en la cara. En otras palabras, las partes faciales de los rostros generados mediante GANs pueden carecer de coherencia o naturalidad al

compararlas con rostros reales. Se señaló que estas anomalías en la disposición de las partes faciales podrían ser evidentes al examinar las ubicaciones de los puntos de referencia faciales, como las puntas de los ojos, la nariz y la boca, los cuales pueden ser detectados de manera efectiva mediante modelos especializados.

Dentro de los trabajos que adoptan un enfoque de aprendizaje profundo, destacan iniciativas como en el trabajo de Xue et al. [79], donde se aborda el análisis de características residuales en imágenes, fusionando este enfoque con el análisis de características fisiológicas, como la estimación de la forma de la pupila y la comparación del color del iris. Asimismo, se resalta el trabajo de Ju et al. [75], el cual implementa un diseño que amalgama información de toda la imagen y características de múltiples secciones seleccionadas mediante un "módulo de selección de parches".

Otro trabajo significativo en este ámbito es el de Li et al. [72], que emplea una red neuronal convolucional multicanal y técnicas de aumento de datos. Esta red acepta datos preprocesados y utiliza mapas de atención para detectar características faciales falsas. El enfoque consta de dos etapas de clasificación: la primera implica el entrenamiento de la red para distinguir entre rostros reales y falsos generados por GANs, mientras que la segunda, denominada etapa de una sola clase, se centra en el entrenamiento para detectar exclusivamente rostros falsos generados, basándose en la distancia entre las características de estos y las de los reales.

Adicionalmente, se destaca el uso recurrente de la red *Xception* en los trabajos [67], [70], [71], [80], la cual ha sido modificada o mejorada. Esto se debe a su capacidad para reducir el número de parámetros y cálculos, manteniendo al mismo tiempo el rendimiento. Esta característica convierte a la red *Xception* en una elección equilibrada en términos de costo computacional y robustez del modelo.

Entre los conjuntos de datos más utilizados en el ámbito de las imágenes naturales se encuentran Flickr-Faces-HQ [3] y *CelebFaces Attributes Dataset* [94], este último con su variante de mayor calidad, CelebA-HQ [95]. Por otro lado, en el dominio de las imágenes generadas artificialmente, sobresalen conjuntos como los generados por *StyleGAN* [3], *StyleGAN2* [2] y *ProGAN / PGGAN* [95].

Además, es importante destacar que, en algunos estudios, no se lleva a cabo experimentación con imágenes que han sido sometidas a técnicas de post-procesamiento, tales como compresión JPEG, desenfoque y ruido gaussiano, redimensionado y realce de contraste, entre otras. La inclusión de estos procesos es fundamental, ya que permite evaluar la robustez del modelo en condiciones no controladas, acercándose así a escenarios más cercanos a la realidad.

Dentro de las métricas utilizadas para evaluar los modelos, se destacan la exactitud, la precisión (P), la tasa de verdaderos positivos y el área bajo la curva *ROC*.

La detección de rostros falsos se puede mejorar en varias áreas. En primer lugar, es necesario abordar el desafío de detectar rostros *GAN* generados localmente; es decir, donde solo se genera una pequeña parte de la imagen. Gran parte de los métodos actuales se centran en caras generadas enteras y pueden no ser eficaces en estos escenarios [67]. En segundo lugar, la falta de interpretabilidad y capacidad de generalización de los métodos basados en el aprendizaje profundo es una limitación. Estos métodos logran una alta precisión, pero tienen dificultades para explicar sus resultados de detección y generalizarlos a través de diferentes métodos de síntesis [18], [74], [80]. Además, confiar en los atributos semánticos de la imagen, como analizar las simetrías e inconsistencias entre los ojos, puede mejorar la detección de rostros generados por *GAN* [17], [18], [19], [79], [96]. Por último, la incorporación de enfoques multifacéticos, como la combinación de la detección del iris con los sistemas de detección de morfología existentes, puede mejorar la precisión de la detección de las morfologías faciales generadas por el *GAN* [17], [71], [72], [75], [79].

Adicionalmente, es mencionado en el trabajo de Guo et al. [18] la idea de explorar la integración de plataformas online para la detección, lo que implica utilizar enfoques ligeros que se adapten a entornos de procesamiento en línea. También, sería beneficioso basarse en más características de los rostros, como la boca o nariz y/o la simetría en el rostro, para fortalecer la detección [17]. Pasquini et al., [16], Li et al., [72] y Guo et al., [66] mencionan considerar distintas redes generadoras de rostros falsos, con principios de funcionamiento distintos a los más explorados dentro del estado del arte, como los generados por *StyleGAN3* [93] y *EG3D* [97] ampliaría la diversidad de posibles manipulaciones de imágenes. Además, mejorar la interpretabilidad de los resultados para los sistemas de aprendizaje profundo, como mediante la localización de zonas de interés, puede aumentar la confianza en los resultados de detección [80]. Finalmente, la experimentación en escenarios prácticos, utilizando muestras "reales" de las redes sociales, sería fundamental para evaluar la eficacia de los métodos de detección en situaciones del mundo real y ajustarlos según sea necesario [66].

En los artículos revisados se observa la capacidad de generalización de los modelos de detección. Específicamente, se destaca la importancia de desarrollar modelos que no solo sean eficientes con las imágenes para las que fueron entrenados, sino que también mantengan cierta robustez ante imágenes generadas por tecnologías distintas a las originales. Esta discusión se hace especialmente relevante en trabajos como los de Guo et al. [45] y Chen et al. [57], donde se explora la detección de imágenes falsas creadas por redes diferentes a aquellas en las que se

entrenaron, ejemplificando con modelos entrenados en imágenes de *StyleGAN2* y probados con imágenes de *PGGAN*, *StyleGAN*, y *StyleGAN3*.

Adicionalmente, se nota una tendencia a realizar pruebas con imágenes sometidas a técnicas de post-procesamiento en un número significativo de trabajos, específicamente en los presentados en [45], [46], [48], [50], [52], [55], [59], [60], [76], [78]. Esto subraya la importancia de considerar cómo el post-procesamiento puede afectar la capacidad de detección de los modelos.

Por otro lado, se identificaron investigaciones que amplían el espectro de análisis a otros tipos de *deepfakes* más allá de los rostros, destacando los trabajos [48], [51], [57]. Los trabajos de Hulzebosch y Guo et al. [52], [98] ponen de relieve no solo la detección de imágenes generadas artificialmente sino también la capacidad humana para diferenciar entre imágenes generadas artificialmente y las naturales, es decir, aquellas pertenecientes a humanos reales, ampliando el enfoque del problema.

El trabajo de Guo et al. [98] introduce una plataforma para evaluar cómo las personas determinan si un rostro ha sido generado artificialmente. Dicho trabajo está organizado en tres etapas, en la primera etapa, los participantes intentan identificar los rostros sin recibir ninguna información adicional. En la segunda, se les proporciona información relevante que podría ayudarles en su evaluación, como detalles sobre la forma de las pupilas. Finalmente, en la tercera etapa, aplican este conocimiento para hacer su evaluación final.

En cuanto a los enfoques metodológicos, los trabajos de Wang y Fu et al. [51] y [46] son destacables por aplicar filtros espaciales, como pasa banda y pasa alta, con el objetivo de minimizar el ruido en la información de la imagen. Asimismo, el estudio de Lu et al. [61], publicado en la segunda mitad de 2023, introduce el uso de *bosques jerárquicos multi-escala en cascada*, analizando las características semánticas, de frecuencia y biológicas en el rostro. Este presenta dos variantes: "*Híbrido*" y "*Divide-and-conquer*", diferenciándose por la adición de capas totalmente conectadas y la evaluación cruzada, respectivamente.

Finalmente, la resolución de las imágenes se presenta como un factor relevante en la detección, con estudios como los de Gragnaniello y Li et al. [82] y [61] que investigan la efectividad de los modelos tanto en imágenes de alta resolución (1024x1024 píxeles) como de baja resolución (256x256 píxeles). Este aspecto destaca la necesidad de adaptar y optimizar los sistemas de detección para operar eficazmente a diferentes escalas de imagen.

Capítulo IV:

Conjuntos de datos

El presente capítulo describe los conjuntos de datos empleados en la experimentación para el desarrollo del modelo de diferenciación de rostros naturales y generados por GAN. Se detallan las características de los conjuntos de datos de rostros naturales y artificiales utilizados.

4.1. Conjuntos de datos de rostros naturales

a) Celebrity Faces Attributes High Quality

El conjunto de datos *CelebA-HQ* es una versión de alta calidad del ya conocido *CelebA* (*CelebFaces Attributes Dataset*) [95]. El conjunto originalmente incluye más de 200,000 imágenes de rostros de celebridades, cada una con 40 anotaciones de atributos y 5 puntos de referencia de localización de características faciales. Estas imágenes, sin embargo, varían en términos de calidad y resolución, lo que puede ser un desafío para aplicaciones que requieren imágenes de alta resolución, como la generación de imágenes con GANs.

Para abordar esta necesidad, T. Karras et al. [95] introdujeron *CelebA-HQ* en 2017 como una versión mejorada del conjunto de datos *CelebA* para su uso en la investigación de GANs, específicamente en su trabajo con *ProGAN*. El proceso de creación de *CelebA-HQ* involucró varios pasos de preprocesamiento para mejorar la calidad de las imágenes, incluyendo técnicas de limpieza de datos y alineación de rostros. En la Figura 4.1 se muestran ejemplos de las imágenes del conjunto.

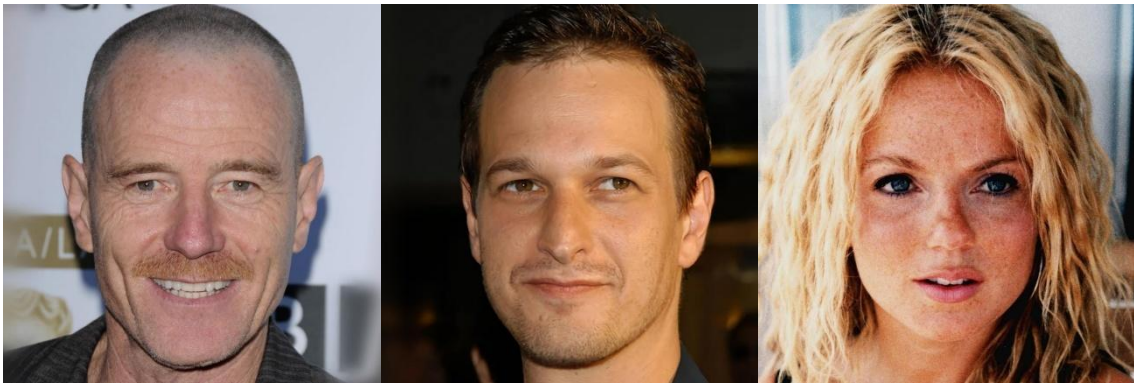


Figura 4.1. Imágenes obtenidas del conjunto CelebA-HQ [95].

CelebA-HQ consta de 30,000 imágenes de alta calidad de 1024x1024 píxeles. La disponibilidad de *CelebA-HQ* ha impulsado numerosos avances en el campo de la visión por computadora y el aprendizaje profundo, facilitando el desarrollo y la evaluación de modelos generativos que pueden crear imágenes de alta resolución.

b) Flickr-Faces-HQ Dataset

Es una colección de imágenes de alta calidad de rostros humanos, introducida en conjunto con la publicación de *StyleGAN*, un tipo avanzado de red generativa adversaria, en 2018. A diferencia de *CelebA-HQ*, que se derivó del conjunto de datos *CelebA* existente, *FFHQ* fue recopilado desde cero para proporcionar una diversidad aún mayor en términos de edad, etnia, y accesorios (como gafas, sombreros, etc.), así como para incluir imágenes con diferentes fondos y condiciones de iluminación.

FFHQ consta de 70,000 imágenes de 1024x1024 píxeles, cada una seleccionada para asegurar la alta calidad y diversidad del conjunto. Las imágenes fueron obtenidas de *Flickr*, una plataforma en línea para compartir fotos, y luego procesadas para uniformar el formato y la resolución. La Figura 4.2 muestra una imagen del conjunto.



Figura 4.2. Imagen obtenida del conjunto FFHQ [3].

El desarrollo de *FFHQ* fue motivado por la necesidad de superar algunas de las limitaciones observadas en conjuntos de datos anteriores, como *CelebA* y *CelebA-HQ*, particularmente en términos de diversidad y calidad de las imágenes. Al incluir una gama más amplia de variabilidad en las imágenes, *FFHQ* permite a los investigadores y desarrolladores entrenar modelos que son capaces de comprender y generar rostros humanos con un grado de detalle y realismo alto.

La introducción de *FFHQ* ha tenido un impacto considerable en el campo de la visión por computadora y el aprendizaje profundo, especialmente en el desarrollo de modelos generativos como *StyleGAN* y sus sucesores. La existencia de *FFHQ* ha facilitado avances en la investigación, permitiendo a la comunidad científica explorar y superar desafíos en la generación de imágenes realistas y en la comprensión de la variabilidad humana a través de la inteligencia artificial.

c) Large-scale Scene Understanding

El conjunto tiene como objetivo ofrecer una referencia alternativa para la clasificación y análisis de escenas a gran escala. *LSUN* [99] abarca 10 categorías distintas de escenas que incluyen escenarios como comedores, dormitorios, iglesias al aire libre, entre otros. En la Figura 4.3 se muestra un ejemplo del conjunto.



Figura 4.3. Ejemplo del conjunto LSUN [99].

Cada categoría en el conjunto de datos de entrenamiento está conformada por una cantidad de imágenes cuyo número varía considerablemente, desde aproximadamente 120,000 hasta 3,000,000 de imágenes por categoría, lo que representa una diversidad considerable en términos de contenido visual y variabilidad de contexto.

d) Conjunto de datos: Animal Faces-HQ

Es un conjunto de imágenes con resolución de 512×512, de tres categorías de animales: gatos, perros y vida silvestre, con aproximadamente 5000 imágenes por cada categoría [100], ejemplos del conjunto se observan en la Figura 4.4. Este conjunto tiene como objetivo abordar problemas en la traducción de imágenes de una categoría a otra, referente a la generación de imágenes.



Figura 4.4. Ejemplos del conjunto AFHQ [100].

En 2021, el conjunto fue actualizado, siendo *AFHQv2*, con algunos cambios en el procesamiento de imágenes. Entre las modificaciones, se usó el filtro *Lanczos* para redimensionar en lugar del método de "vecino más cercano", lo cual mejora la calidad visual al reducir el *aliasing*. Las imágenes fueron guardadas en formato PNG para evitar artefactos de compresión, lo cual aumenta el tamaño de los archivos, pero mantiene la calidad.

4.2. Conjuntos de datos de rostros generados

A continuación, se detallan los conjuntos de datos compuestos por rostros creados mediante inteligencia artificial, acompañados de una descripción breve de la arquitectura utilizada para generarlos y el conjunto de datos utilizado para el entrenamiento del modelo generador utilizado.

a) Conjunto generado por ProGAN

Conjunto de datos de rostros generados por la arquitectura *PROGAN* [95] siendo entrenada con el conjunto de datos *CelebA-HQ*.

PROGAN es una arquitectura de *GAN* introducida en 2017. Esta arquitectura se caracteriza por su capacidad de generar imágenes de alta resolución y calidad de manera más eficiente y estable en comparación con las técnicas de GANs antecesoras.

La clave detrás de *PROGAN* es su enfoque progresivo para el entrenamiento de la red. En lugar de intentar generar imágenes de alta resolución desde el principio, la red comienza generando imágenes de baja resolución. A medida que el entrenamiento avanza, se van agregando capas a la red, lo que permite incrementar gradualmente la resolución de las imágenes generadas. Este enfoque reduce significativamente los problemas de estabilidad durante el entrenamiento y permite que la red aprenda a generar detalles más finos de manera incremental.

Es importante destacar que, la calidad de las imágenes generadas se evalúa utilizando el *FID* (*Frechet Inception Distance*), una métrica que cuantifica la distancia entre las distribuciones de características de las imágenes generadas y las imágenes reales. Un *FID* más cercano a 0 indica una menor distancia entre las imágenes generadas y las reales, lo que se considera una medida de calidad superior. En el caso de *PROGAN*, la red alcanza un *FID* de 7.79, lo que indica un rendimiento bueno en la generación de imágenes de alta calidad.

El resultado del entrenamiento de *PROGAN* en *CelebA-HQ* fue la generación de un conjunto de datos de 100,000 imágenes de rostros en resolución 1024x1024. Un ejemplo del contenido del conjunto se muestra en la Figura 4.5.



Figura 4.5. Imagen obtenida del conjunto ProGAN [95].

b) Conjunto generado por StyleGAN

La arquitectura de *StyleGAN* [3], desarrollada por investigadores de *NVIDIA*, representa un hito importante en el campo de la generación de imágenes mediante el uso de GANs. *StyleGAN* introduce varios conceptos innovadores que permiten la generación de imágenes de alta calidad con un nivel de control sin precedentes sobre las características generadas, obteniendo un valor *FID* de 4.40. A continuación, se detallan los aspectos clave de la arquitectura *StyleGAN* y las innovaciones que implementa:

- **Descomposición del Estilo:** Una de las principales innovaciones de *StyleGAN* es su capacidad para descomponer y manipular el "estilo" de las imágenes generadas en diferentes niveles de detalle. Esto se logra mediante un mapeo del espacio latente a un espacio de estilos intermedio, permitiendo controlar aspectos específicos de la imagen, como la pose, la forma facial, las características, el cabello, y la iluminación, de manera independiente.
- **Mapeo del Espacio Latente:** Se introduce un mapeo no lineal del espacio latente a un espacio de estilos intermedio a través de una red de mapeo. Esto permite un control más fino sobre las variaciones de las imágenes generadas y ayuda a desentrañar los factores de variación de las imágenes de manera más efectiva.
- **Normalización Adaptativa de Instancias:** se utiliza en *StyleGAN* para ajustar el estilo de las imágenes en cada nivel de la jerarquía de generación. AdaIN modula las características de la imagen en cada escala, permitiendo la

manipulación de estilos específicos en diferentes resoluciones, desde características globales como la forma de la cabeza hasta detalles finos como la textura de la piel.

- **Inyección de Ruido:** *StyleGAN* introduce ruido aleatorio en varios puntos de la red generadora, lo que permite la creación de detalles finos y texturas realistas en las imágenes generadas. Este ruido se añade de manera controlada para influir en características específicas de la imagen, como la presencia de pecas o la textura del cabello, añadiendo así variabilidad y realismo a las imágenes.

Posterior a su desarrollo, se entrenó *StyleGAN* utilizando el conjunto de datos FFHQ, logrando generar 100,000 imágenes de rostros con una resolución de 1024x1024, un ejemplo se aprecia en la Figura 4.6. La elección del conjunto *FFHQ* para el entrenamiento fue debido a su variedad en términos de edad, etnias y expresiones faciales.



Figura 4.6. Imagen obtenida del conjunto *StyleGAN* [3].

c) Conjunto generado por *StyleGAN2*

Se enfoca en corregir artefactos característicos y mejorar la calidad de los resultados en comparación con su predecesor, *StyleGAN*. Una característica distintiva de *StyleGAN* es su arquitectura de generador no convencional, que introduce un código latente intermedio para controlar estilos en la red de síntesis a través de normalización adaptativa de instancias

StyleGAN2 [2] mejora la metodología, logrando una proyección casi perfecta de las imágenes generadas de vuelta a las entradas del generador, a diferencia de las imágenes reales que muestran diferencias claras. Indicando avances significativos en la calidad de imagen y la atribución de imágenes generadas a su fuente. Además, *StyleGAN2* logra un rendimiento de entrenamiento mejorado.

Los artefactos son defectos visuales no deseados en las imágenes generadas por GANs. *StyleGAN* exhibe artefactos característicos en forma de gota marcados con círculos rojos, como se muestra en la Figura 4.7. Es posible que la gota no sea obvia en la imagen final, pero está presente en los mapas de características intermedias del generador.



Figura 4.7. Artefactos en forma de gota en imágenes generadas por StyleGAN [2].

Para solucionar estos problemas, *StyleGAN2* presenta una mejora significativa al introducir una arquitectura de generador de dos ramas, lo que reduce drásticamente la presencia de artefactos. Una imagen del conjunto se muestra en la Figura 4.8.



Figura 4.8. Imagen obtenida del conjunto StyleGAN2 [2].

La implementación y los modelos, así como un conjunto de datos de 100,000 imágenes de rostros con resolución de 1024x1024 generadas por *StyleGAN2*, entrenado con el conjunto de datos *FFHQ*, están disponibles públicamente, facilitando la experimentación.

d) Conjunto generado por StyleGAN2-ADA

En 2020, *StyleGAN2-ADA* [101] mejora su versión predecesora de varias maneras, especialmente en contextos donde los conjuntos de entrenamiento son limitados. La introducción de la adaptación discriminatoria de aumentos permite una mejora

significativa en la calidad y diversidad de las imágenes generadas incluso con conjuntos de datos pequeños.

Los métodos que no emplean aumentos, como *StyleGAN2*, muestran una degradación notable en la calidad y diversidad de las imágenes generadas cuando el tamaño del conjunto de entrenamiento disminuye, generalmente produciendo resultados de baja calidad con menos de 30,000 imágenes de entrenamiento. En contraste, con *ADA*, la degradación es mucho más suave, y los resultados siguen siendo razonables incluso con un conjunto de entrenamiento de 5k imágenes.

Se destaca que, sin *ADA*, la convergencia se ve obstaculizada por el sobreajuste del discriminador, lo que lleva a una calidad de imagen inferior para el *StyleGAN2* original, especialmente en conjuntos de datos más pequeños. Esto se muestra claramente en los resultados generados para varios conjuntos de datos, donde *ADA* logra resultados excelentes en todos los casos.

El conjunto de datos contiene 263 imágenes de 1024x1024 píxeles de resolución, utilizando *FFHQ* para su entrenamiento, véase Figura 4.9.



Figura 4.9. Imagen obtenida del conjunto *StyleGAN2-ADA* [101].

e) Conjunto generado por *StyleGAN3*

Mejora significativamente a *StyleGAN2*, a través de la supresión adecuada del *aliasing*, logra una mejor jerarquía en la generación de imágenes y videos, con representaciones internas que permiten una correcta adhesión de detalles a las superficies. Introduce cambios menores para prevenir la filtración de información no deseada, logrando que las redes resultantes sean equivariantes a la traslación y rotación incluso a escalas subpixel [93].

StyleGAN3 introduce dos variantes principales: *StyleGAN3-T* y *StyleGAN3-R*. Ambas versiones se centran en mejorar la equivarianza y eliminar artefactos, pero se diferencian en sus capacidades específicas:

- *StyleGAN3-T* está diseñado generar imágenes con una consistencia mejorada cuando se aplican transformaciones espaciales, como mover la imagen. Este modelo mantiene la competitividad con *StyleGAN2* en términos de *FID*, que es una medida de cuán realistas son las imágenes generadas en comparación con un conjunto de datos real.
- *StyleGAN3-R*, por otro lado, no solo mantiene las capacidades del modelo anterior y añade equivarianza de rotación. Esto significa que *StyleGAN3-R* puede manejar de manera efectiva las rotaciones de imágenes, asegurando que los detalles y las texturas de las imágenes generadas permanezcan consistentes incluso cuando la imagen se gira.

Se tienen 2 pares de subconjuntos (véase Figura 4.10) los cuales fueron entrenados con *FFHQ*, y tienen una resolución de 1024x1024 píxeles, los subconjuntos *U*, son imágenes en las cuales el rostro presente no se encuentra alineado en el centro de ésta, es decir, se cuenta con:

- *StyleGAN3-R* (226 imágenes).
- *StyleGAN3-T* (334 imágenes).
- *StyleGAN3-U-R* (164 imágenes).
- *StyleGAN3-U-T* (143 imágenes).



Figura 4.10. Imágenes obtenidas del conjunto *StyleGAN3*. *StyleGAN3-T* (superior izquierda). *StyleGAN3-T* (superior derecha). *StyleGAN3-U-R* (inferior izquierda). *StyleGAN3-U-T* (inferior derecha) [93].

En la Tabla 4.1 se presenta un resumen de los conjuntos de datos empleados en la presente investigación. Esta tabla clasifica los conjuntos según contengan rostros naturales o generados artificialmente, e indica la cantidad de imágenes que los componen. Es pertinente señalar que el asterisco denota aquellos conjuntos en los que no todas las imágenes se encuentran disponibles. Cabe destacar la inclusión del conjunto *LSUN*, reconocido por su gran escala y diversidad de escenas. Sin embargo, debido a que este conjunto data de 2015, la descarga completa y fiable desde su repositorio oficial ha presentado ciertas dificultades, lo cual constituye una limitación en su accesibilidad.

Tabla 4.1. Resumen de los conjuntos de datos.

Conjunto	Clase	Número de imágenes
FFHQ	Real	70,000
CelebA-HQ	Real	30,000
LSUN	Real	+150,000*
AFHQ	Real	15,000
ProGAN	GAN	100,000
StyleGAN	GAN	100,000
StyleGAN2	GAN	100,000
StyleGAN2-ADA	GAN	263
StyleGAN3	GAN	867

* significa que no todas las imágenes se encuentran disponibles.

La caracterización y el manejo de estos conjuntos de datos fueron fundamentales para la fase experimental, permitiendo entrenar y validar el modelo con una amplia gama de ejemplos de rostros reales y sintéticos, a pesar de las particularidades observadas en la obtención de ciertos recursos como el conjunto *LSUN*.

Capítulo V:

Desarrollo del sistema

En este capítulo se presenta detalladamente el proceso de experimentación llevado a cabo para desarrollar y evaluar el modelo propuesto para la diferenciación de rostros naturales y generados artificialmente. Se describen las distintas fases de experimentación, los escenarios de evaluación utilizados para medir el rendimiento del modelo y el análisis de los resultados obtenidos.

5.1. Entorno de desarrollo

Hardware

El hardware utilizado para la experimentación consistió en los siguientes componentes:

- GPU: RTX 4060ti con 16GB de VRAM.
- Memoria RAM: 32GB DDR5.
- Procesador: Intel Core i5 12400.

Software

El entorno de software se basó en el sistema operativo *Windows 11*. Para la codificación y desarrollo, se empleó *Visual Studio Code*. El lenguaje de programación principal fue *Python*, utilizando la versión 3.8.10. Para facilitar el desarrollo del modelo de aprendizaje profundo y el procesamiento de datos, se integraron diversas bibliotecas de *Python*. Los módulos de *Python* clave utilizados y sus versiones específicas se listan en la siguiente tabla:

Tabla 5.1. Módulos de Python utilizados.

Modulos pip	Versión
torch	1.9.0
numpy	1.19.2
scikit-learn	0.24.2
tqdm	4.62.0
pandas	1.3.0
timm	0.4.12

Este entorno de desarrollo proporcionó los recursos necesarios para la implementación, entrenamiento y la evaluación del modelo de diferenciación de rostros, permitiendo el manejo eficiente de los conjuntos de datos extensos y la ejecución de los complejos algoritmos de aprendizaje profundo.

5.2. Experimentación

El objetivo de esta sección es detallar el proceso de experimentación realizado para desarrollar y optimizar un modelo capaz de diferenciar entre rostros naturales y rostros generados artificialmente, evaluando diversas arquitecturas, técnicas de aumento de datos y configuraciones de parámetros para lograr el mejor rendimiento en la clasificación respecto a los conjuntos *FFHQ*, *StyleGAN* y *StyleGAN2*, siendo estos seleccionados previamente como objetivo de clasificación.

Adicionalmente, se realizan pruebas con los conjuntos *CelebA-HQ*, *ProGAN*, *StyleGAN2-ADA* y *StyleGAN3*, siendo un total de dos conjuntos naturales y cinco conjuntos generados por modelos GAN. El objetivo es obtener de manera cuantitativa el mejor rendimiento en el reconocimiento de los rostros generados, para posteriormente, realizar ajustes enfocados a mejorar el rendimiento del sistema, no solo enfocándose en los conjuntos objetivo. Las pruebas realizadas son las siguientes:

- a) Entrenamiento inicial.
- b) Aumento de datos mediante técnicas tradicionales.
- c) Experimentación con distintas arquitecturas utilizando un conjunto reducido.
- d) Elección y experimentación con la mejor arquitectura.
- e) Mejoras en los parámetros del modelo final.
 - *Dropout*, optimizador y programador de tasa de aprendizaje.
 - Epocas de calentamiento.
 - Cambios en la función de pérdida.
 - Congelamiento de etapas.

- Cambios en el conjunto de entrenamiento.
- Aumento del tamaño de la imagen de entrada.
- f) Evaluación de nuevas técnicas de aumento de datos.
 - Aumento de datos basado en *Deep Dream*.
 - Aumento de datos mediante deformación de imágenes.
 - Ajustes de volumen en los conjuntos y el aumento de datos.
 - Evaluación de la eficacia de los métodos de aumento de datos propuestos, en el ámbito de clasificación de rostros generados artificialmente.

Estas pruebas llevan a obtener un modelo final que ofrece una buena robustez respecto a las pruebas con los conjuntos previamente mencionados.

a) Entrenamiento inicial

Se utilizó el modelo pre-entrenado 'vit_base_patch16_224' tal como se proporciona en *timm* [4], utilizando los módulos *timm* y *PyTorch*. Este modelo trabaja con imágenes de tamaño 224x224 píxeles y las divide en 16 parches, siguiendo el modelo base propuesto por Dosovitskiy et al. [28].

- Se empleó el optimizador *AdamW*, conocido por su eficacia en la regularización y manejo del decaimiento de pesos, lo que ayuda a prevenir el sobreajuste y mejora la convergencia del modelo.
- Se seleccionó una tasa de aprendizaje de 0.001 y se aplicó un decaimiento de pesos de $1e-5$ para equilibrar la velocidad de aprendizaje, la regularización y la estabilidad durante el entrenamiento de modelos complejos.
- Además, se incorporó el uso de la función *GradScaler* para manejar la precisión mixta durante el entrenamiento, optimizando el rendimiento sin sacrificar la precisión numérica.

El modelo se entrenó con un conjunto de datos de 40,000 imágenes, compuesto por 20,000 imágenes de la clase real del conjunto *FFHQ* y 20,000 imágenes de la clase GAN, divididas en 10,000 generadas por *StyleGAN* y 10,000 por *StyleGAN2* (véase Tabla 5.2). El conjunto de datos es dividido en 80% para entrenamiento y 20% para validación.

Tabla 5.2. Conjunto de datos de entrenamiento y de la validación.

Conjuntos	Número de imágenes
Clase real (FFHQ [3])	20,000
Clase GAN (StyleGAN [3] / StyleGAN2 [2])	10,000 / 10,000

El entrenamiento se llevó a cabo durante un máximo de 100 épocas, con una detención o parada temprana; es decir, si después de 5 épocas, no hay mejora en la pérdida de validación, se interrumpe el proceso. Esto se lleva a cabo para evitar el sobreajuste y reducir el tiempo de cómputo innecesario.

A continuación, en la Tabla 5.3, se muestran los subconjuntos (del conjunto de prueba), así como el número de imágenes que contiene cada uno.

Es importante señalar que los dos últimos conjuntos: *StyleGAN2-ADA* y *StyleGAN3*, no alcanzan las 1,000 imágenes debido a la limitación en la cantidad de imágenes disponibles en sus respectivos repositorios.

Tabla 5.3 Conjunto de datos para las pruebas.

Conjuntos	Número de imágenes
FFHQ[3]	1,000
CelebA-HQ [95]	1,000
StyleGAN [3]	1,000
ProGAN [95]	1,000
StyleGAN2 [2]	1,000
StyleGAN2-ADA [101]	263
StyleGAN3 [93]	867

En la Figura 5.1 se muestran las gráficas de su entrenamiento y de la validación. En la Tabla 5.4 se muestran los resultados de las pruebas posteriores, donde se observa una exactitud del 76.1% en el conjunto *StyleGAN* y del 99.6% en el conjunto *StyleGAN2*. Adicionalmente, los resultados en los conjuntos *ProGAN*, *StyleGAN2-ADA* y *StyleGAN3* son menores al 10%, lo que indica que el modelo obtenido no presenta una buena capacidad de generalización.

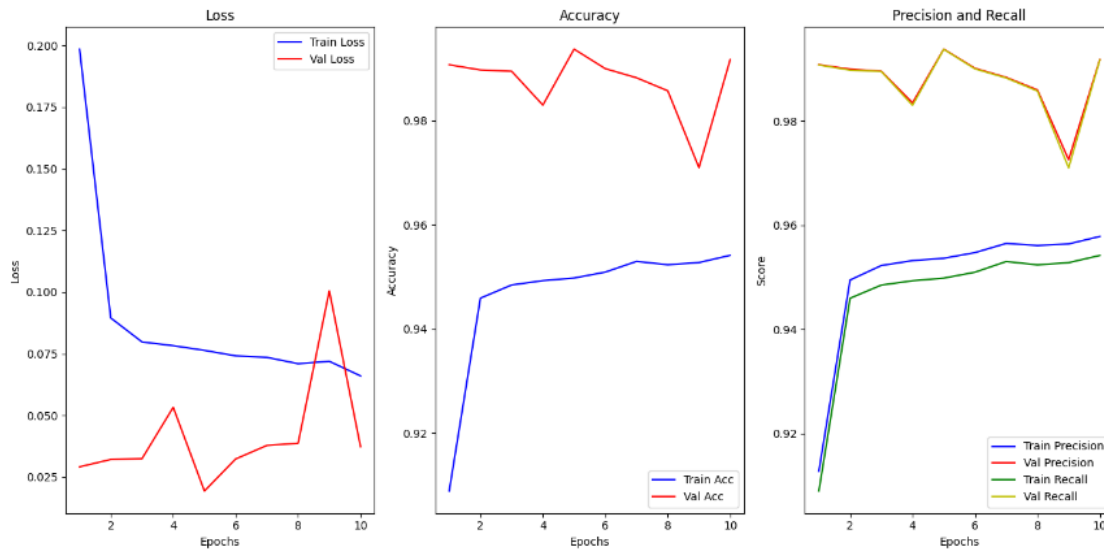


Figura 5.1 Gráficas de entrenamiento de la experimentación inicial.

Tabla 5.4. Exactitud (%) en las pruebas posteriores de la experimentación inicial.

FFHQ	CelebA-HQ	StyleGAN	ProGAN	StyleGAN2	StyleGAN2-ADA	StyleGAN3
99.8	99.9	76.1	0	99.6	6.4	1.3

b) Aumento de datos mediante técnicas tradicionales

Se añadieron transformaciones de imágenes para un aumento de datos en línea (véase Tabla 5.5). Con el fin de incrementar la diversidad del conjunto de datos sin aumentar su tamaño físico y con ello ayudar al modelo a generalizar mejor y reducir el riesgo de sobreajuste.

Tabla 5.5. Técnicas de aumento de datos utilizadas.

Técnica	Descripción	Valores
<i>RandomHorizontalFlip</i>	Invierte horizontalmente la imagen. Se aplica aleatoriamente con probabilidad p.	p=0.5
<i>ColorJitter</i>	Modifica aleatoriamente el brillo, contraste, saturación y tono de la imagen.	brightness=0.3, contrast=0.3, saturation=0.3, hue=0.3
<i>RandomAffine</i>	Aplica transformaciones como rotación, traslación y deformación. Se aplica aleatoriamente con probabilidad p.	degrees=30, translate=(0.25, 0.25), shear=20
<i>RandomGrayscale</i>	Convierte la imagen a escala de grises. Se aplica aleatoriamente con probabilidad p.	p=0.3
<i>RandomPerspective</i>	Aplica una transformación de perspectiva aleatoria para distorsionar la imagen.	distortion_scale=0.5, p=0.5
<i>RandomErasing</i>	Borra aleatoriamente una región rectangular de la imagen.	p=0.5, scale=(0.02, 0.1), ratio=(0.3, 3.3), value=0
<i>RandomResizedCrop</i>	Recorta y redimensiona aleatoriamente una parte de la imagen.	size=224, interpolation='lanczos', scale=(0.4, 0.5), ratio=(3/3, 4/4)
<i>RandomCrop</i>	Realiza un recorte aleatorio de la imagen.	size=80
<i>RandomAdjustSharpness</i>	Ajusta aleatoriamente la nitidez de la imagen.	sharpness_factor=1.5
<i>RandomAutocontrast</i>	Aplica un ajuste de contraste. Se aplica aleatoriamente con probabilidad p.	p=0.5
<i>RandomEqualize</i>	Ecualiza el histograma de la imagen. Se aplica aleatoriamente con probabilidad p.	p=0.5
<i>GaussianBlur</i>	Aplica un desenfoque gaussiano a la imagen.	kernel_size=3, sigma=(0.1, 3)

Además, se utilizó *ReduceLROnPlateau*, como programador de tasa de aprendizaje. Este ajusta automáticamente la tasa de aprendizaje reduciéndola cuando la métrica monitorizada, como la pérdida de validación, deja de mejorar. Con el cambio se obtienen mejoras con los conjuntos *ProGAN*, *StyleGAN2-ADA* y *StyleGAN3*, los resultados de estas modificaciones se observan en la Tabla 5.6, el comportamiento de las métricas de entrenamiento y de la validación se muestra en la Figura 5.2.

Tabla 5.6. Exactitud (%) en las pruebas posteriores de la experimentación con aumento de datos.

FFHQ	CelebA-HQ	StyleGAN	ProGAN	StyleGAN2	StyleGAN2-ADA	StyleGAN3
99.3	99.7	86.5	1.9	100	17.11	4.0

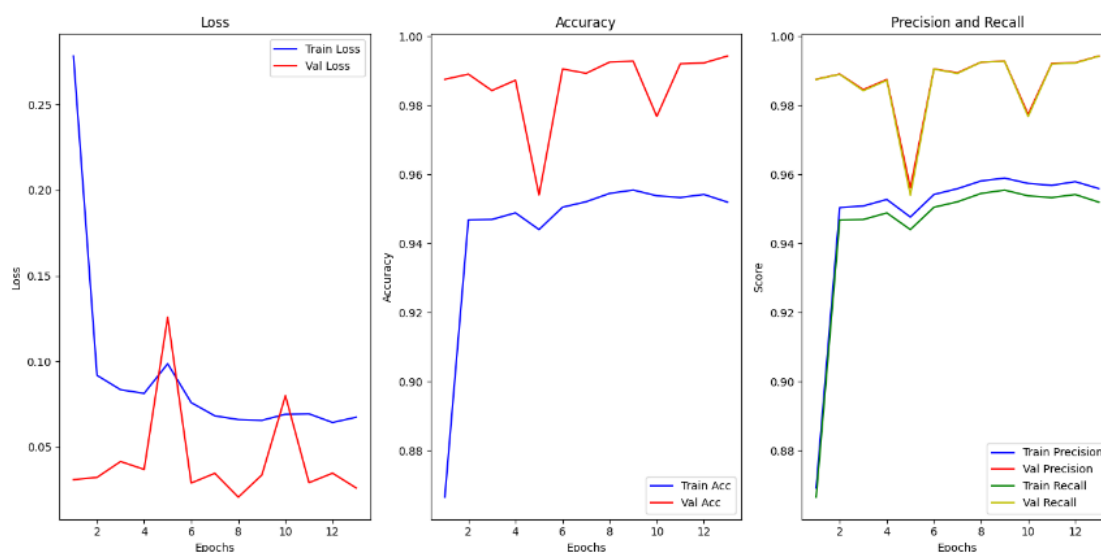


Figura 5.2 Gráficas de entrenamiento de la experimentación con aumento de datos.

c) Experimentación con distintas arquitecturas utilizando un conjunto reducido

El objetivo fue evaluar el rendimiento de un modelo pequeño, para ello se seleccionó el '*vit_tiny_patch16_224*' [28], con el aumento de datos activo (véase Tabla 5.7 y Figura 5.3). En donde no se percibió un cambio considerable.

Tabla 5.7. Exactitud (%) en las pruebas posteriores utilizando el modelo '*vit_tiny_patch16_224*'.

FFHQ	CelebA-HQ	StyleGAN	ProGAN	StyleGAN2	StyleGAN2-ADA	StyleGAN3
99.7	99.8	80.1	0.2	99.7	14.0	4.2

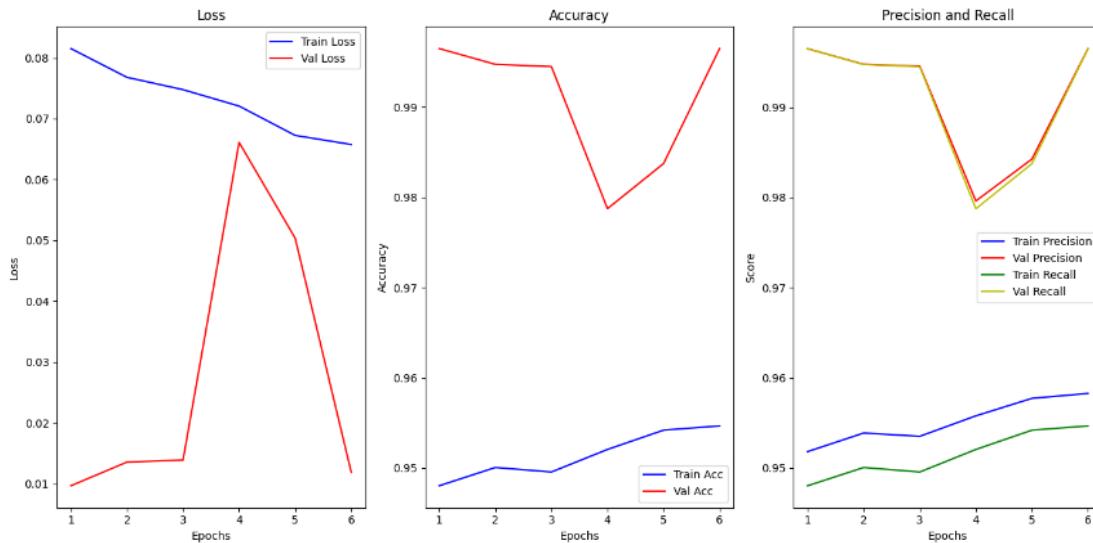


Figura 5.3 Gráficas de entrenamiento utilizando el modelo 'vit_tiny_patch16_224'.

Se redujo el tamaño del conjunto de datos de entrenamiento en tres valores, pasando de 40,000 imágenes a: 20,000, 10,000 y finalmente 5,000, con la distribución detallada en la Tabla 5.8, esto debido a que se observó que no ocurría un cambio notable en el rendimiento de los modelos.

Tabla 5.8. Conjunto de datos de entrenamiento y de la validación con 5,000 muestras.

Conjuntos	Número de imágenes
Clase real (FFHQ [3])	2,500
Clase GAN (StyleGAN [3] / StyleGAN2 [2])	1,250 / 1,250

En la Tabla 5.9 y Figura 5.4, se muestran los resultados del entrenamiento con el conjunto de 5,000 muestras.

Tabla 5.9. Exactitud (%) en las pruebas posteriores con el conjunto de 5,000 muestras.

FFHQ	CelebA-HQ	StyleGAN	ProGAN	StyleGAN2	StyleGAN2-ADA	StyleGAN3
97.4	98.8	68.9	0.6	99.3	22.8	10.3

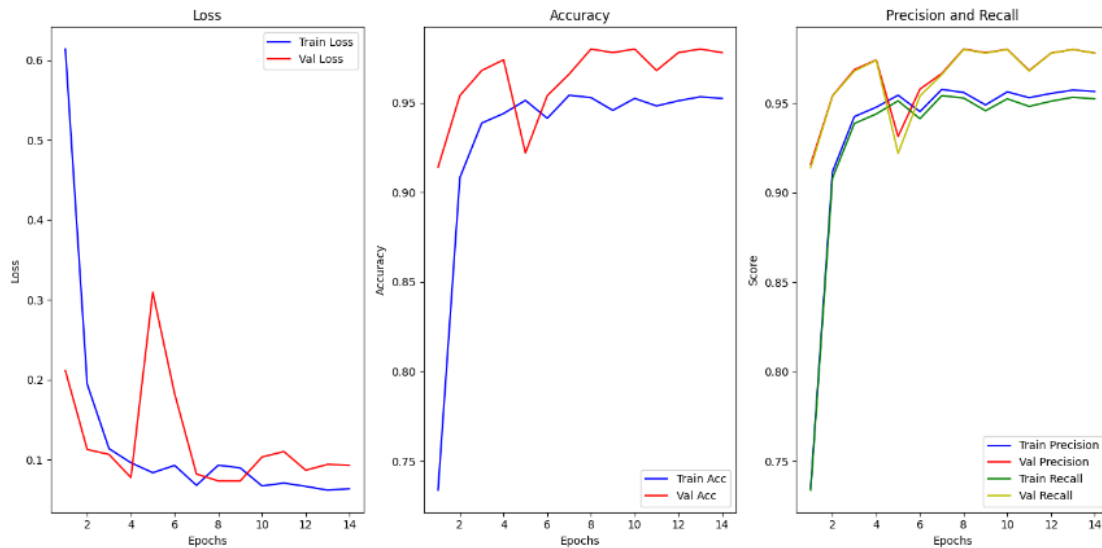


Figura 5.4 Gráficas de entrenamiento con el conjunto de 5,000 muestras.

De acuerdo con los resultados obtenidos, a pesar de la diferencia entre el tamaño de los conjuntos de entrada, esto no repercute en un cambio notable en el rendimiento de los modelos y los valores de las métricas se mantienen casi iguales. Finalmente, se realizaron pruebas con un conjunto de entrenamiento de 2,500 datos, pero en este caso, el rendimiento disminuyó más de 10 puntos porcentuales.

En resumen, con el fin de agilizar los procesos de entrenamiento y pruebas correspondientes a la experimentación con distintas arquitecturas, se decidió mantener el conjunto con 5,000 datos como mínimo, ya que como se dijo, no se observaron cambios significativos en las pruebas comparado con conjuntos más grandes y hay un cambio sustancial en el tiempo.

Experimentación con distintas arquitecturas

Timm [4] ofrece la posibilidad de trabajar con alrededor de 390 modelos, provenientes de aproximadamente 15 familias de ViTs. Los modelos se seleccionaron tratando de utilizar un modelo por familia, dando prioridad a modelos catalogados como pequeños en sus respectivos trabajos, con sufijos como *tiny* o *small*, con el objetivo de reducir el tiempo de entrenamiento de los mismos. Asimismo, en caso de que el trabajo reportase un modelo como destacable frente a otras variantes, es utilizado para obtener un mejor resultado, como es el caso de 'crossvit_18_dagger_240' [102].

Los resultados obtenidos con el conjunto de evaluación, con cada uno de los modelos implementados fueron similares a los obtenidos previamente. En la Tabla

5.10 se observa la exactitud obtenida en el conjunto de prueba con los modelos utilizados:

Tabla 5.10. Exactitud (%) en los modelos utilizados.

Modelo	FFHQ	CelebA-HQ	StyleGAN	ProGAN	StyleGAN 2	StyleGAN 2-ADA	StyleGAN 3
convit_tiny.fb_in1k [103]	98.2	97.4	59.8	1.7	98.8	14.4	8
crossvit_18_dagger_240.in1k [102]	97.9	97.4	73.6	4	99.3	24.7	17.5
davit_tiny.msft_in1k [104]	99.7	99.7	88.7	6.4	99.9	33.4	7.3
efficientvit_m5.r224_in1k [105]	94.5	94.4	63.7	2.7	98.7	21.2	12.4
fastvit_ma36.apple_dist_in1k [106]	100	100	81.50	4.40	99.9	23.9	2.7
gcvit_base.in1k [107]	99.8	99.6	95.8	14.4	99.9	46.3	20.4
mobilevitv2_075.cv_nets_in1k [108]	97.3	98.3	72.2	3.6	98.1	21.6	10
nextvit_small.bd_in1k [109]	99.6	99.7	84.7	0.9	99.8	26.2	5.4
vitamin_small_224.datacomp1b_clip [110]	98	98.3	62.5	1.7	97.7	17.1	6
'mvitv2_tiny.fb_in1k' [29]	99.2	99.3	89.8	23.7	98.9	44.11	19.3

Obteniendo hasta 33.4%, 44.11% y 46.3% de exactitud en el conjunto *StyleGAN2-ADA* y 19.3% y 20.4% de exactitud en el conjunto *StyleGAN3* con los dos últimos. Sin embargo, la decisión de seguir experimentando con la arquitectura *mvitv2* se justificó por la disponibilidad de diversos modelos más grandes y con mejoras aplicadas, lo que ofrecía un mayor potencial para explorar la escalabilidad y adaptación del modelo en comparación con *gcvit*, que solo contaba con una variante base.

d) Elección y experimentación con la mejor arquitectura

Con el modelo *'mvitv2_tiny.fb_in1k'*, se observa una mejora significativa en los resultados, especialmente en los conjuntos *ProGAN*, *StyleGAN2-ADA* y *StyleGAN3*. Este modelo mostró el mejor desempeño durante la experimentación, por lo que se decidió continuar utilizando esta arquitectura y sus variantes. Los modelos *'mvitv2_base.fb_in1k'*, *'mvitv2_base_cls.fb_inw21k'* y *'mvitv2_large_cls.fb_inw21k'*, son arquitecturas más grandes y, en el caso de los dos últimos, cuentan con ajuste

fino [29]. Lo anterior, motiva a continuar con la experimentación con esa familia de arquitecturas, en la Tabla 5.11 se muestran los resultados obtenidos tras realizar el entrenamiento, validación y pruebas con esta arquitectura.

Tabla 5.11. Exactitud (%) en los modelos MViTv2.

Modelo	FFHQ	CelebA-HQ	StyleGAN	ProGAN	StyleGAN 2	StyleGAN 2-ADA	StyleGAN 3
'mvitv2_base .fb_in1k'	99.7	99.9	95.2	37.6	100	56.2	24.9
'mvitv2_base _cls. fb_inw21k'	99.5	99.7	99.1	84.3	99.8	80.6	50.4
'mvitv2_large _cls.fb_inw 21k'	99.6	99.8	95.2	21.4	99.7	42.9	15.4

Dado los resultados anteriores, se decidió trabajar con el modelo **'mvitv2_base_cls.fb_inw21k'**, siendo uno de los modelos que cuentan con ajuste fino en su entrenamiento inicial [29] y habiendo obtenido los resultados más altos. Una vez elegido el modelo pre entrenado con el que se trabajará, se decidió continuar con el conjunto de 40,000 muestras, para los entrenamientos posteriores, para contar con más ejemplos.

e) Mejoras en los parámetros del modelo final

Con el objetivo de conocer la influencia e importancia de los diferentes parámetros con los cuales se configura el modelo final, se presentan de manera parcial la modificación y la evaluación de cada uno de ellos.

Dropout, optimizador y programador de tasa de aprendizaje

La arquitectura elegida cuenta con una capa de *Dropout* con valores de 0 antes de la capa completamente conectada encargada de elegir la clase. Esta capa se activó asignándole un valor de 0.25, es decir aplicando 25% de *Dropout*.

A su vez, se decidió cambiar el optimizador, de *AdamW* a *AdamP*, debido a que presenta un mejor desempeño en tareas de clasificación según la documentación de *timm* [111]. Ello se puede ver en la Figura 5.5, en donde se comparan los optimizadores *SGD*, *Adam*, *AdamW*, *Nadam*, *Radam*, *AdamP* y *Lookahead_Adam*. El eje X de la gráfica representa el número de épocas, y el eje Y la métrica de pérdida.

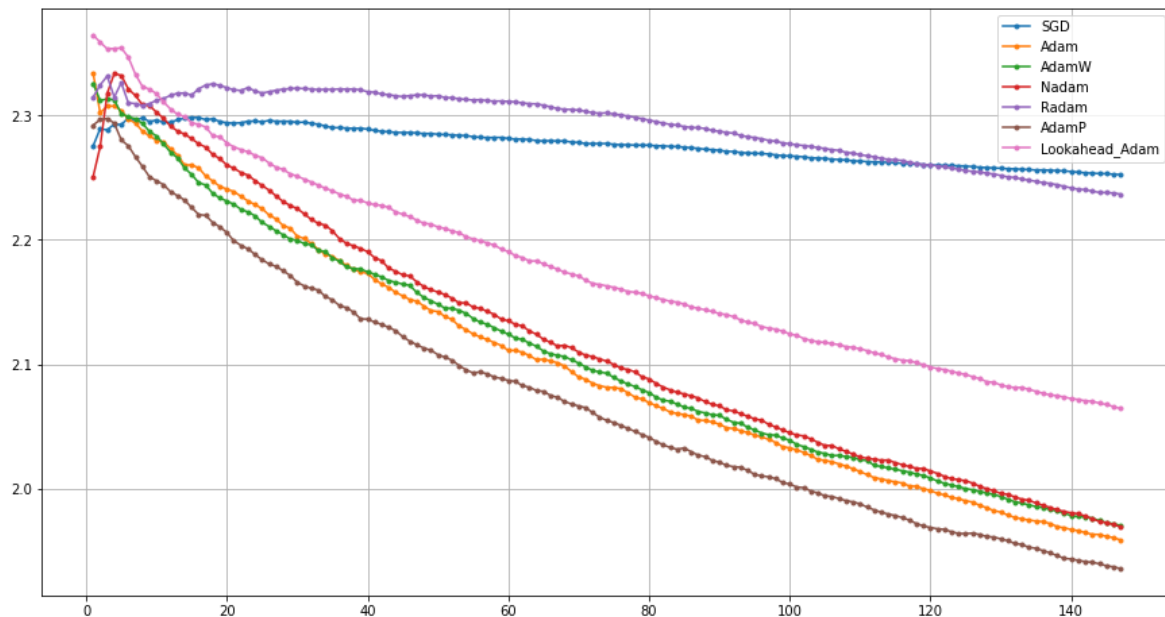


Figura 5.5 Comparación de optimizadores en tareas de clasificación [111].

De igual manera se cambió el ajuste de la tasa de aprendizaje de *ReduceLROnPlateau* a *CosineAnnealingLR* para lograr una disminución suave y cíclica, favoreciendo una mejor convergencia y evitando el sobreajuste. Esto permite una exploración más eficiente del espacio de parámetros, manteniendo la tasa de aprendizaje activa en distintos niveles.

En la Figura 5.6 se muestran las gráficas resultantes del entrenamiento y de la validación, en la Tabla 5.12 resultados de las pruebas posteriores.

La pérdida de entrenamiento y de la validación muestran un descenso más controlado y menos fluctuaciones en las últimas épocas. Esto indica un ajuste más estable del modelo, de igual manera la exactitud converge mejor, reduciendo la brecha entre ambas curvas.

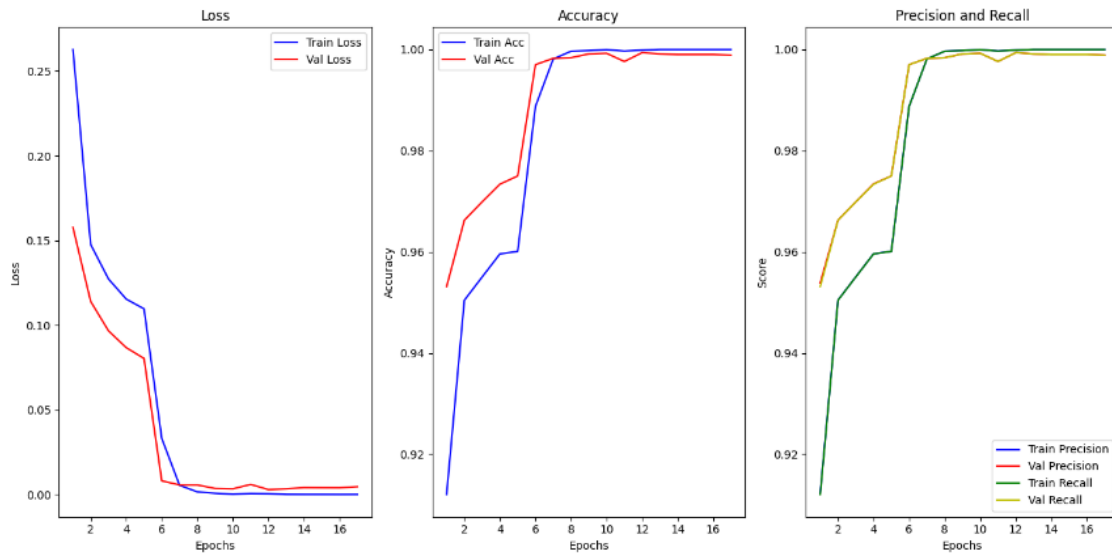


Figura 5.6 Gráficas de entrenamiento al aplicar cambios en el dropout, optimizador y ajuste de tasa de aprendizaje.

Tabla 5.12. Exactitud (%) en las pruebas posteriores al aplicar cambios en el dropout, optimizador y ajuste de tasa de aprendizaje.

FFHQ	CelebA-HQ	StyleGAN	ProGAN	StyleGAN2	StyleGAN2-ADA	StyleGAN3
99.8	100	99.4	80.1	100	82.8	46.6

Épocas de calentamiento

Se implementaron épocas de calentamiento como una estrategia para mejorar la estabilidad durante el entrenamiento del modelo, utilizando dos épocas en donde se congela la mayor parte de los pesos pre entrenados del modelo. En esta fase, se utilizó una tasa de aprendizaje baja, concretamente de $1e-05$, que aumenta a $1e-03$ al comenzar las épocas posteriores al calentamiento. Este proceso permitió al modelo ajustarse gradualmente a los patrones de los datos, evitando fluctuaciones abruptas y facilitando una convergencia más estable y eficiente en etapas posteriores del entrenamiento. Como resultado, se observa una mejora en el comportamiento del modelo durante el aprendizaje y, el conjunto de validación presenta una mejor respuesta que el conjunto de entrenamiento, ver la Figura 5.7. El comportamiento se mantuvo sin cambios significativos en los resultados de las pruebas posteriores (véase Tabla 5.13).

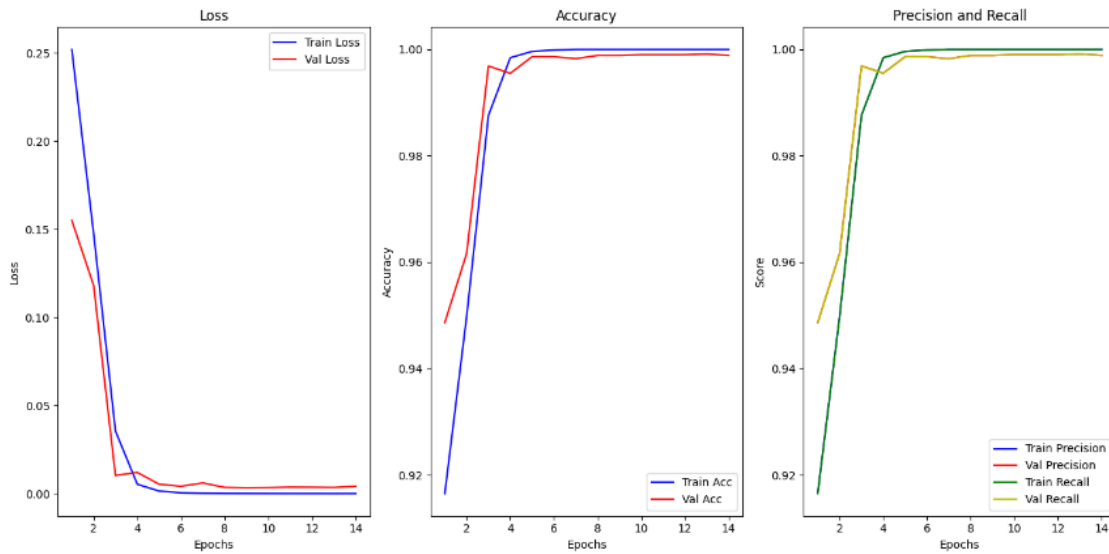


Figura 5.7 Gráficas de entrenamiento al implementar épocas de calentamiento.

Tabla 5.13. Exactitud (%) en las pruebas posteriores al implementar épocas de calentamiento.

FFHQ	CelebA-HQ	StyleGAN	ProGAN	StyleGAN2	StyleGAN2-ADA	StyleGAN3
99.8	100	99.2	78.8	100	79.4	46.8

Cambios en la función de pérdida

Se decidió cambiar la función de pérdida de *CrossEntropy* a *Focal Loss* para evaluarla. El cambio resultó en una mejora significativa en la exactitud del modelo (véase Tabla 5.14 y Figura 5.8), especialmente en los conjuntos de datos *ProGAN* y *StyleGAN2-ADA*.

La función *Focal Loss* permitió al modelo enfocarse más en las muestras difíciles de clasificar, lo que contribuyó a un incremento en la precisión global sin afectar la estabilidad del entrenamiento. Obteniendo mejoras principalmente en los conjuntos de *ProGAN*, *StyleGAN2-ADA* y *StyleGAN3*, de 4.3, 5.7 y 4.4 puntos porcentuales respectivamente.

Tabla 5.14. Exactitud (%) en las pruebas posteriores al utilizar 'Focal Loss' como función de pérdida.

FFHQ	CelebA-HQ	StyleGAN	ProGAN	StyleGAN2	StyleGAN2-ADA	StyleGAN3
99.7	99.9	99.3	82.2	100	84.0	49.9

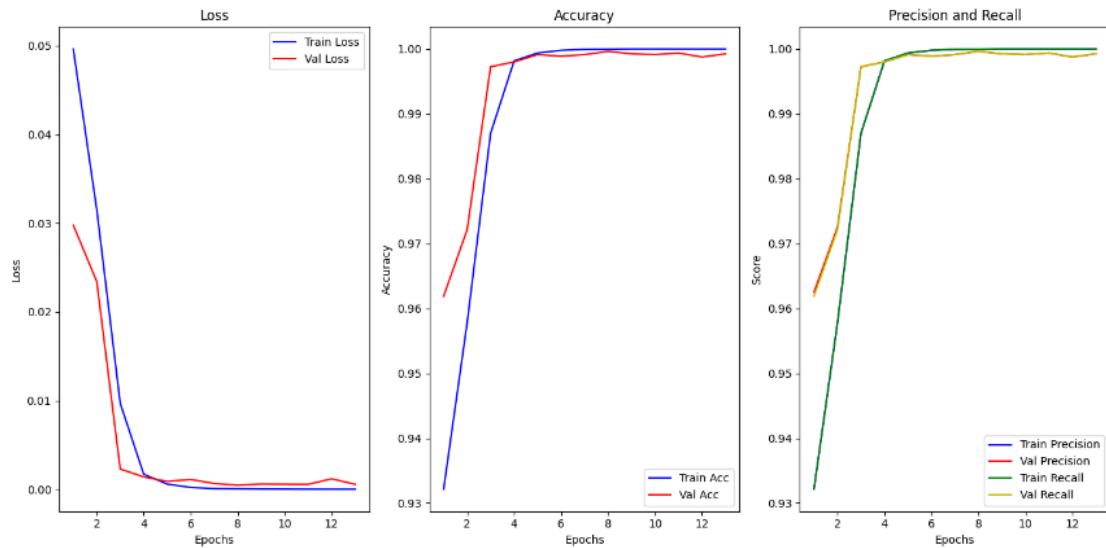


Figura 5.8 Gráficas de entrenamiento al utilizar 'Focal Loss' como función de pérdida.

Congelamiento de etapas

Se evaluó la configuración de congelar toda la red durante el entrenamiento, excepto las etapas 3 y 4, lo que implica mantener congeladas todas las etapas desde el inicio del entrenamiento hasta la etapa 2. Este ajuste tuvo un impacto negativo en el rendimiento del modelo y su comportamiento durante el entrenamiento (véase Figura 5.9), resultando en una disminución notable de la exactitud, como se observa en la Tabla 5.15.

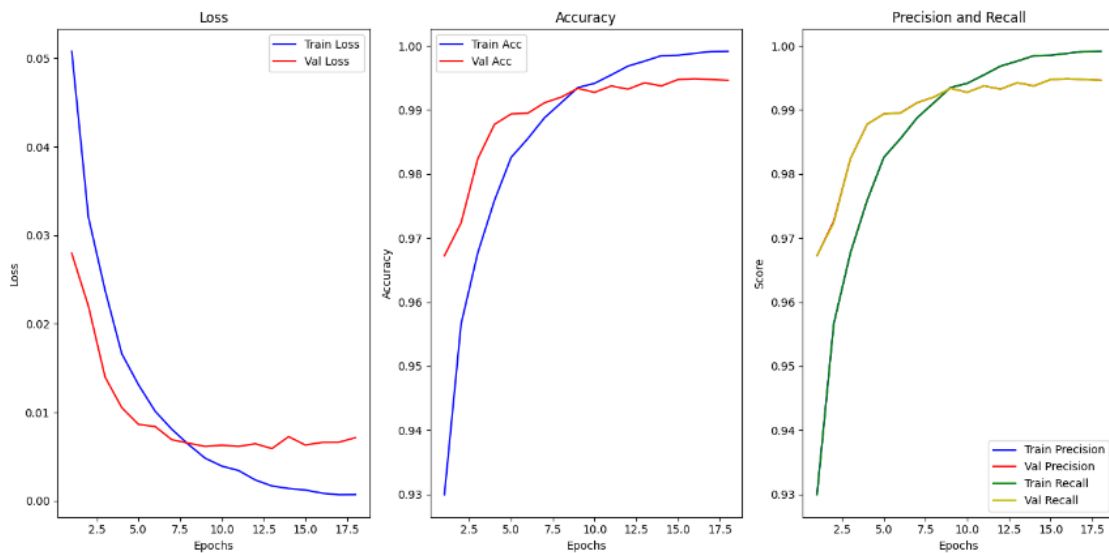


Figura 5.9 Gráficas de entrenamiento al congelar los bloques 0, 1 y 2 de la arquitectura.

Tabla 5.15. Exactitud (%) en las pruebas posteriores al congelar los bloques 0, 1 y 2 de la arquitectura.

FFHQ	CelebA-HQ	StyleGAN	ProGAN	StyleGAN2	StyleGAN2-ADA	StyleGAN3
99.2	99.4	91.3	53.8	99.9	41.4	27.1

Posteriormente, se optó por congelar únicamente la etapa 0, es decir las capas anteriores a la etapa 1, con la esperanza de mantener el balance entre la extracción de características y la eficiencia del modelo. Sin embargo, aunque este enfoque también redujo la exactitud, como se ve en la Tabla 5.16, el desempeño fue mejor en comparación con la configuración anterior. A pesar de ello, ambas configuraciones fueron descartadas, ya que no lograron mejorar el rendimiento general del modelo.

Tabla 5.16. Exactitud (%) en las pruebas posteriores al congelar únicamente el bloque 0 de la arquitectura.

FFHQ	CelebA-HQ	StyleGAN	ProGAN	StyleGAN2	StyleGAN2-ADA	StyleGAN3
99.8	100	98.8	70.9	99.9	70.7	39.5

Cambios en el conjunto de entrenamiento

Se modificó el conjunto de entrenamiento (ver Tabla 5.17). Para las clases reales, se mantuvo el conjunto *FFHQ*. Con respecto a las clases generadas artificialmente, *StyleGAN2* se conservó y se reemplazó *StyleGAN* por *ProGAN*. Este cambio se realizó debido a que, el conjunto *ProGAN* contiene rostros generados artificialmente que no logran parecer completamente humanos, debido a que es una de las primeras arquitecturas GAN en ser desarrolladas, con una capacidad de generación realista inferior a las arquitecturas actuales.

Tabla 5.17 Modificaciones en el conjunto de entrenamiento.

Conjuntos	Número de imágenes
Clase real (FFHQ [3])	20,000
Clase GAN (ProGAN [95] / StyleGAN2 [2])	10,000 / 10,000

Este ajuste produjo una mejora en las pruebas posteriores, especialmente con los conjuntos de datos *ProGAN* y *StyleGAN2-ADA*, donde se observa una mejora en la precisión de aproximadamente 7 puntos porcentuales (véase Tabla 5.18). Sin embargo, en el conjunto de *StyleGAN*, se registró una disminución en el rendimiento, reflejando una baja de, aproximadamente 7 puntos porcentuales.

Tabla 5.18. Exactitud (%) en las pruebas posteriores al congelar únicamente el bloque 0 de la arquitectura.

FFHQ	CelebA-HQ	StyleGAN	ProGAN	StyleGAN2	StyleGAN2-ADA	StyleGAN3
99.9	100	92.1	100	99.8	82.9	43.9

Aumento del tamaño de la imagen de entrada

Se decidió aumentar el tamaño de las imágenes de entrada de 224x224 píxeles a 384x384 píxeles con el propósito de introducir mayor cantidad de información visual al modelo, lo que mejora la extracción de características.

Este cambio resultó en mejoras significativas en las pruebas posteriores al entrenamiento, como se observa en la Tabla 5.19 y Figura 5.10, especialmente en los conjuntos de datos *StyleGAN*, *StyleGAN2-ADA* y *StyleGAN3*.

Como resultado de este cambio se observa que el aumento en la resolución permite al modelo distinguir con mayor precisión entre los detalles sutiles que caracterizan a los rostros generados por las diferentes arquitecturas, incrementando así la exactitud en la clasificación.

Tabla 5.19. Exactitud (%) en las pruebas posteriores después de ajustar el tamaño de entrada del modelo.

FFHQ	CelebA-HQ	StyleGAN	ProGAN	StyleGAN2	StyleGAN2-ADA	StyleGAN3
100	100	98.6	100	100	89.7	48.1

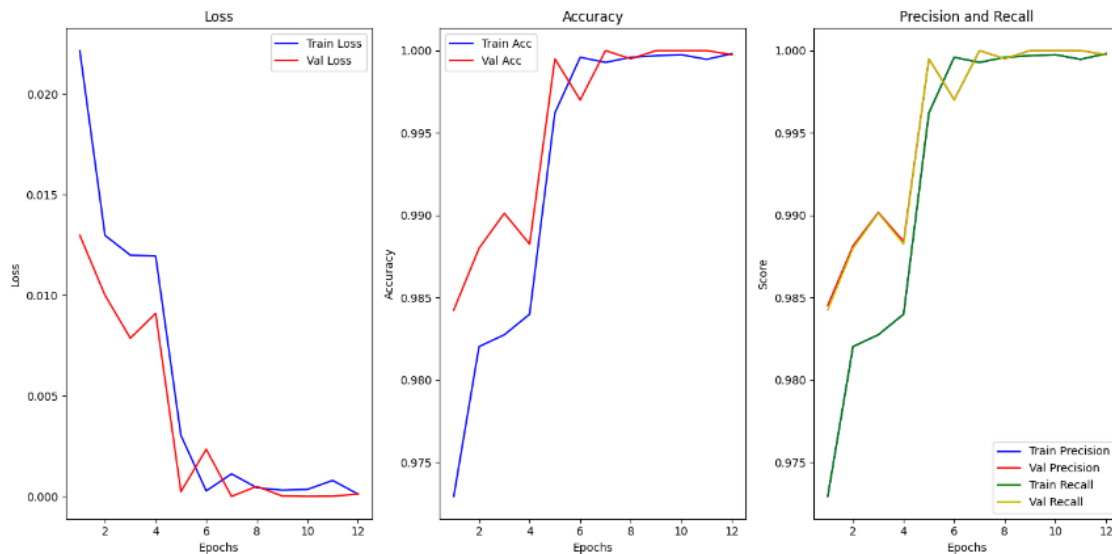


Figura 5.10 Gráficas de entrenamiento después de ajustar el tamaño de entrada del modelo.

f) Evaluación de nuevas técnicas de aumento de datos

Aumento de datos basado en *Deep Dream*

Al revisar las imágenes clasificadas incorrectamente, se observó que algunas contenían artefactos en las texturas. Estos artefactos presentaban características similares a las generadas por el efecto *Deep Dream*, las cuales alteran la textura y añaden patrones visuales peculiares a las imágenes originales. En la Figura 5.11 se presentan algunos ejemplos de estas imágenes con artefactos, lo que motivó la decisión de explorar el uso de técnicas de aumento de datos que incorporen este tipo de transformaciones.



Figura 5.11 Ejemplos de imágenes con acercamiento en artefactos.

En el trabajo de Karras et al. [93], se muestran ejemplos de representaciones internas del modelo (ver Figura 5.12) donde se pueden apreciar patrones visuales similares a los producidos por *Deep Dream*. Se concluyó que ciertos aspectos de estas representaciones internas permanecen en la imagen generada final. Estos hallazgos sugieren que la presencia de tales artefactos es una característica

inherente a las imágenes generadas por redes de tipo GAN, lo que dificulta la discriminación efectiva entre imágenes naturales y generadas.

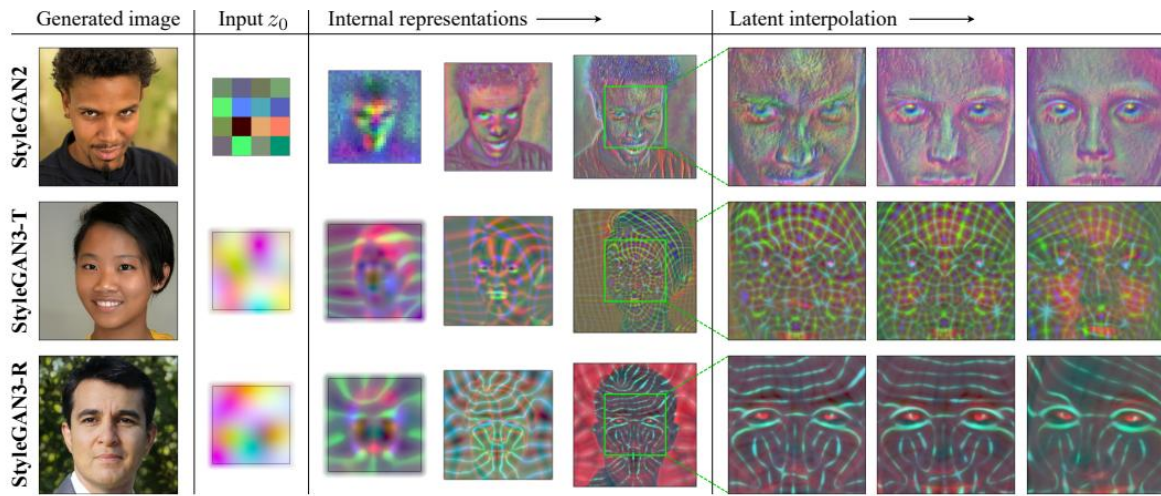


Figura 5.12 Ejemplos de representaciones internas de los modelos [93].

Considerando esta evidencia, se optó por incluir un aumento de datos fuera de línea, aplicando el efecto de *Deep Dream* a imágenes del conjunto de rostros naturales. Para implementar esta técnica, se utilizó la red VGG16, ajustando las imágenes en cinco iteraciones mediante la optimización de la norma de la salida del modelo. Se empleó una tasa de aprendizaje de 0.001 y se utilizó el gradiente para modificar la imagen, definiendo el porcentaje de imagen a cambiar en el 45% de ésta.

Además, se aplicó una máscara elíptica aleatoria para definir regiones específicas en las que se implementaría el efecto de *Deep Dream*, asegurando que solo se alteren las áreas seleccionadas, las cuales son aleatorias.

Se implementó el efecto sobre imágenes naturales y las imágenes resultantes fueron asignadas a la clase generada. El mantener las versiones originales en el conjunto natural permite al modelo detectar diferencias sutiles entre un rostro natural y uno que aparenta ser natural, pero con artefactos evidentes, como se observa en la Figura 5.13. Adicionalmente, las tasas de aprendizaje en las épocas de calentamiento y posteriores al calentamiento fueron reducidas a 1e-06 y 1e-04 respectivamente, con el fin de estabilizar el proceso de entrenamiento, permitiendo que el modelo ajuste de manera más precisa sus parámetros. Este ajuste busca evitar que se produzcan oscilaciones indeseadas en la convergencia, asegurando que las diferencias sutiles entre ambas clases sean capturadas efectivamente.



Figura 5.13 Cambios en la imagen al aplicar filtro Deep Dream en una zona.

La inclusión de imágenes con artefactos similares a los de *Deep Dream* en el conjunto de datos, permite al modelo reconocer y discriminar de manera más efectiva los patrones presentes en los rostros generados, optimizando así la capacidad del modelo para distinguir entre imágenes reales y generadas artificialmente con características más sutiles.

Aumento de datos mediante deformación de imágenes

Examinando las muestras clasificadas incorrectamente en los casos anteriores se encontró que existen ejemplos en donde, ciertas zonas de la imagen muestran efectos similares a deformaciones, como se observa en la Figura 5.14.



Figura 5.14 Ejemplos de imágenes generadas artificialmente con deformaciones.

Por lo que, de manera similar a lo realizado con el aumento de datos basado con *Deep Dream*, se añadió otra técnica de aumento de datos fuera de línea, en la cual se define una elipse que distorsiona la imagen aplicando un efecto sinusoidal en un ángulo al azar, con la posibilidad de seleccionar cuan fuerte será el efecto en la imagen resultante. Se definió el factor de distorsión en un rango entre 10% y 15%, y un porcentaje de cambio en la imagen del 10%. Un ejemplo se puede ver en la Figura 5.15.



Figura 5.15 Cambios en la imagen al aplicar deformación en una zona.

g) Ajustes de volumen en los conjuntos y el aumento de datos

Siguiendo el procedimiento descrito en los repositorios oficiales [112], [113], se generaron imágenes para los conjuntos *StyleGAN2-ADA* y *StyleGAN3*, obteniendo 1,000 muestras por conjunto, con los parámetros de generación por defecto, en el caso de *StyleGAN3*, al tener cuatro variantes que varían en términos de rotación, traslación y alineación del rostro, como se menciona en el marco teórico, obteniendo 250 imágenes por variante para un total de 1,000.

En la Tabla 5.20 se muestra cómo quedó integrado el conjunto de prueba después de estas modificaciones, los conjuntos modificados se muestran en negritas. Esta ampliación del conjunto de datos mejora la fiabilidad de los resultados obtenidos durante la fase de evaluación del modelo, ya que permite una mejor comparación entre los diferentes subconjuntos. Se recalca que este cambio no produce una mejora cuantitativa respecto a las pruebas generadas, sino en la calidad de la evaluación al proporcionar un contexto más amplio y representativo para el análisis.

Tabla 5.20 Modificación del conjunto de prueba.

Conjuntos	Número de imágenes
FFHQ[3]	1,000
CelebA-HQ [95]	1,000
StyleGAN [3]	1,000
ProGAN [95]	1,000
StyleGAN2 [2]	1,000
StyleGAN2-ADA [101]	1,000
StyleGAN3 [93]	1,000

En el conjunto de entrenamiento se incluyen ejemplos de *CelebA-HQ* en la clase real, y en la clase generada se añaden ejemplos con efecto *Deep Dream (DD)*, con deformación de imágenes (*W*), y con ambas, en la Tabla 5.21 se muestran los datos del conjunto de entrenamiento utilizado.

Tabla 5.21. Conjunto de entrenamiento al realizar ajuste del volumen de aumento de datos.

Conjuntos	Número de imágenes
Clase real (FFHQ [3] / CelebA-HQ [95])	10,000 / 10,000
Clase GAN (ProGAN [95] / StyleGAN2 [2] / Real-DD / Real-W / Real-DDW)	9,500 / 9,500 / 333 / 333 / 334

También se realizaron cambios en la condición para un paro temprano en el entrenamiento, añadiendo un umbral de cambio mínimo para la activación de éste. Es decir, la métrica observada, en este caso la pérdida en la validación debe bajar más del valor del umbral para ser considerada como una época con mejora, el valor del umbral fue definido como 0.005.

Capítulo VI:

Resultados

El presente capítulo se dedica a la exposición detallada de los resultados obtenidos a partir de la metodología de investigación propuesta. Los resultados se desglosarán en función de los objetivos específicos planteados en este trabajo, incluyendo métricas cuantitativas que validan la eficacia del modelo, así como un análisis cualitativo de su desempeño.

6.1. Resumen del modelo final

La solución final utiliza como base el modelo *'mvitv2_base_cls.fb_inw21k'*. A continuación, se resume la configuración y los principales parámetros aplicados. Primeramente, se detalla el conjunto de datos (véase Tabla 6.1) con el que fue entrenado, se utilizó una división de 80% de datos para el entrenamiento y 20% para la validación del modelo.

Tabla 6.1. Conjunto de datos de entrenamiento, incluyendo el aumento de datos fuera de línea.

Conjuntos	Número de imágenes
Clase real (FFHQ [3] / CelebA-HQ [95])	10,000 / 10,000
Clase GAN (ProGAN [95] / StyleGAN2 [2] / Real-DD / Real-W / Real-DDW)	9,500 / 9,500 / 333 / 333 / 334

Adicionalmente, en la Tabla 6.2 se presentan nuevamente las técnicas utilizadas para el aumento de datos en línea, junto a una descripción breve y los valores utilizados en la técnica.

Tabla 6.2. Técnicas de aumento de datos utilizadas en línea.

Técnicas	Descripción	Valores
<i>RandomHorizontalFlip</i>	Invierte horizontalmente la imagen. Se aplica aleatoriamente con probabilidad p.	p=0.5
<i>ColorJitter</i>	Modifica aleatoriamente el brillo, contraste, saturación y tono de la imagen.	brightness=0.3, contrast=0.3, saturation=0.3, hue=0.3
<i>RandomAffine</i>	Aplica transformaciones como rotación, traslación y deformación. Se aplica aleatoriamente con probabilidad p.	degrees=30, translate=(0.25, 0.25), shear=20
<i>RandomGrayscale</i>	Convierte la imagen a escala de grises. Se aplica aleatoriamente con probabilidad p.	p=0.3
<i>RandomPerspective</i>	Aplica una transformación de perspectiva aleatoria para distorsionar la imagen.	distortion_scale=0.5, p=0.5
<i>RandomErasing</i>	Borra aleatoriamente una región rectangular de la imagen.	p=0.5, scale=(0.02, 0.1), ratio=(0.3, 3.3), value=0
<i>RandomResizedCrop</i>	Recorta y redimensiona aleatoriamente una parte de la imagen.	size=224, interpolation='lanczos', scale=(0.4, 0.5), ratio=(3/3, 4/4)
<i>RandomCrop</i>	Realiza un recorte aleatorio de la imagen.	size=80
<i>RandomAdjustSharpness</i>	Ajusta aleatoriamente la nitidez de la imagen.	sharpness_factor=1.5
<i>RandomAutocontrast</i>	Aplica un ajuste de contraste. Se aplica aleatoriamente con probabilidad p.	p=0.5
<i>RandomEqualize</i>	Ecualiza el histograma de la imagen. Se aplica aleatoriamente con probabilidad p.	p=0.5
<i>GaussianBlur</i>	Aplica un desenfoque gaussiano a la imagen.	kernel_size=3, sigma=(0.1, 3)

Asimismo, se recopilan los valores finales de los parámetros en los cuales se realizaron cambios a lo largo de la experimentación antes detallada (véase Tabla 6.3). Se hace hincapié en que el valor del umbral de paro temprano especifica que la mejora en la pérdida de validación debe ser mayor a este umbral. El número de épocas máximo para el entrenamiento es definido como 100; sin embargo, el paro temprano se activó en la época 11, definiendo los resultados de la octava época como los mejores.

Tabla 6.3. Parámetros del entrenamiento de la red.

Parámetros	Valores
Tamaño de imagen	384x384 píxeles
Función de pérdida	<i>Focal loss</i>
Decaimiento de pesos	1E-05
<i>Dropout</i>	25%
Algoritmo optimizador	<i>AdamP</i>
Ajuste de tasa de aprendizaje automático	<i>CosineLR</i>

Épocas de calentamiento	2
Tasa de aprendizaje en épocas de calentamiento	1E-06
Tasa de aprendizaje inicial posterior a calentamiento	1E-04
Épocas de paro temprano	3
Umbral de paro temprano	0.005

Se define este modelo como final, mismo con el que se realizarán las pruebas correspondientes a todos los escenarios de evaluación.

6.2. Evaluación de la eficacia de los métodos de aumento de datos propuestos, en el ámbito de clasificación de rostros generados artificialmente

El objetivo de este experimento es verificar que las técnicas de aumento de datos ‘*Deep Dream*’ y deformación de imagen representan una mejora dentro del tema de clasificación entre imágenes reales y GAN.

El procedimiento consiste en realizar pruebas con distintos conjuntos de entrenamiento y luego verificar la robustez del modelo utilizando las imágenes generadas, sin la aplicación de procesos posteriores o condiciones diferentes.

Las condiciones de entrenamiento son las siguientes:

- Sin aumento de datos.
- Aumento de datos típico.
- Aumento de datos típico y *Deep dream*.
- Aumento de datos típico y deformación de imágenes.
- Aumento de datos típico junto a *Deep dream* y deformación de imágenes.
- Aumento con *Deep dream*.
- Aumento con deformación de imágenes.
- Aumento con *Deep dream* y deformación de imágenes.

Resultados

Para este caso, se realizaron 7 entrenamientos nuevos (posteriores al modelo final), en los cuales todos los parámetros y mejoras realizadas durante el desarrollo del proyecto se mantuvieron. El cambio en los nuevos entrenamientos fue únicamente en los conjuntos de datos y el aumento de datos aplicado que fueron de tres tipos.

Los resultados se observan en la Tabla 6.4, en donde “T” se refiere a un aumento de datos típico, utilizado dentro del ámbito del aprendizaje profundo, las técnicas de aumento de datos correspondientes a *T* se encuentran en la Tabla 5.5, “DD” se refiere a la técnica implementada basada en *Deep dream* y “W” se refiere a la técnica de deformación de imágenes.

Tabla 6.4. Resultados en términos de exactitud con las configuraciones probadas.

Configuración	FFHQ	CelebA-HQ	ProGAN	StyleGAN	StyleGAN2	StyleGAN2-ADA	StyleGAN3	Promedio
Sin aumento de datos	99.6	100	94.9	100	100	90.7	66.3	93.07
T	99.7	100	94.3	100	100	89.9	66.6	92.93
DD	100	99.9	98.9	100	100	94.4	86.3	97.07
W	99.7	100	98.2	100	100	95.6	94.8	98.33
T+DD	100	99.9	98.8	100	100	94.6	87.4	97.24
T+W	99.5	99.9	98.4	100	100	95.3	93.2	98.04
DD+W	100	100	99.4	100	100	98.3	61.8	94.21
T+DD+W	99.6	99.9	99.7	100	100	95.7	94.8	98.39

Se observa que, el mejor desempeño promedio de los 7 conjuntos, lo obtuvo la configuración *T+DD+W* con 98.39% de exactitud. Sin embargo, también destaca la configuración *W*, logrando casi el mismo desempeño, con 0.06 puntos porcentuales de diferencia, lo que permite concluir que el aumento de datos con técnicas de “deformación” aportan información al modelo de manera significativa

Se destaca que, respecto a conjuntos más avanzados como *StyleGAN3*, las técnicas de aumento de datos propuestas, en específico *DD* y *W*, contribuyen notablemente, aumentando 19.7 y 28.2 puntos porcentuales en el desempeño frente a las imágenes generadas por el modelo.

6.3. Escenarios de evaluación

Escenario principal

Utilizando el conjunto de prueba, integrado por imágenes diferentes a las consideradas en los conjuntos de entrenamiento y de la validación se obtuvieron las métricas generales del modelo, que se muestran en la Tabla 6.5.

Tabla 6.5. Métricas del modelo final.

Métrica	Valor (%)
Exactitud	98.39
Precisión	99.90
TPR / Sensibilidad	97.84
TNR / Especificidad	99.50
Puntuación F1	98.86
Área bajo la curva ROC	98.80

La Tabla 6.6 destaca los resultados específicos de la exactitud en cada conjunto, dentro de las pruebas. En ellos se observa hasta 94.8% de exactitud en el conjunto de imágenes generadas por el modelo de *StyleGAN3*. Este análisis permite observar la robustez del modelo respecto a la capacidad de generación de algunas de las principales redes actuales, proporcionando una visión clara de cómo el modelo clasifica imágenes generadas por distintas arquitecturas. Esto no solo resalta la efectividad del modelo en identificar patrones en imágenes generadas por *StyleGAN3*, sino que también brinda un marco de referencia para evaluar su desempeño frente a otros generadores. Estos resultados pueden servir como base para futuras mejoras y comparaciones en el ámbito de la clasificación de imágenes generadas artificialmente.

Tabla 6.6. Exactitud del modelo en los conjuntos probados. * Significa que el conjunto está presente en el entrenamiento.

Conjunto	Exactitud (%)
FFHQ*	99.6
CelebA-HQ*	99.9
ProGAN*	98.7
StyleGAN	100
StyleGAN2*	100
StyleGAN2-ADA	95.7
StyleGAN3	94.8

En la Tabla 6.7, se presenta una comparativa respecto a 3 trabajos del estado del arte. En la tabla se observa que los valores de exactitud obtenidos con el modelo final destacan al ser superiores, a excepción de las imágenes generadas por el conjunto *ProGAN*, en donde el trabajo de Xue et al. [79] reporta un aumento en la exactitud de 1.1 puntos porcentuales.

Tabla 6.7. Comparación con trabajos del estado del arte en términos de exactitud.

Conjunto	Modelo obtenido	Wang et al. [17]	Xue et al. [79]	Lu et al. [61]
ProGAN	98.7	93.6	99.8	-
StyleGAN	100	-	98.4	100
StyleGAN2	100	100	99.1	98.7
StyleGAN3	94.8	92.2	-	88.4

Continuando con los resultados del modelo, en la Tabla 6.8, se muestra la matriz de confusión obtenida, destacando que, si bien se redujo la tasa de falsos negativos, es decir, las imágenes *GAN* predichas como reales, persiste un porcentaje mínimo.

Tabla 6.8. Matriz de confusión del modelo.

		Predicho	
		Real	GAN
Verdad	Real	1995	5
	GAN	108	4892

Escenario 2: Análisis de imágenes de rostros generadas mediante modelos GAN, sometidas a operaciones de procesamiento posterior a la generación

La experimentación es abordada en el estado del arte, en trabajos como [45], [46], [50], [52], [55], [59], [60], [76], [78], que buscan demostrar la capacidad del modelo original en diferentes condiciones de imagen.

El objetivo es comprobar la robustez del modelo desarrollado respecto a las imágenes con las siguientes operaciones aplicadas:

- Compresión JPEG de 90%, 80% y 70%.
- Ruido gaussiano con una desviación estándar de la distribución de 5.
- Desenfoque gaussiano con tamaños de kernel 3x3 y 5x5.
- Corrección gamma con valor 1.1.
- Filtro mediana con un kernel de tamaño 3x3.

- Redimensionamiento a tamaño de imagen de 75% y 50% del tamaño de entrada utilizado.

Resultados

Para este escenario, se generaron 10 conjuntos de prueba, utilizando las imágenes del conjunto original, aplicando diferentes operaciones a las imágenes. El tamaño de entrada al modelo es de 384x384 píxeles, por lo que el redimensionamiento aplicado se lleva a cabo considerando dicho tamaño y, no los 1024x1024 del tamaño original. El ruido gaussiano es aplicado con una desviación estándar de valor cinco.

Los resultados obtenidos se muestran en la Tabla 6.9, en donde se observa la robustez del modelo frente a las operaciones consideradas.

Tabla 6.9. Resultados de exactitud en pruebas con imágenes con procesamiento posterior a la generación.

Operación	FFHQ	CelebA-HQ	PG	SG1	SG2	SG2-ADA	SG3
Filtro de mediana 3x3	99.5	99.8	97.9	100	100	95.6	93.3
Desenfoque gaussiano 3x3	99.1	99.7	88.7	99.5	97.8	79.2	71.7
Desenfoque gaussiano 5x5	97.7	99.5	65.7	96.9	81.90	50.9	41.4
Ruido gaussiano	99.7	96	88.7	96.8	96.5	44.2	58.3
Redimensión a 75%	99.4	99.9	95.5	99.7	99.1	85.8	81.9
Redimensión a 50%	100	100	54.7	78.2	53	25.9	24.1
Corrección gamma 1.1	100	100	57.6	75.6	65.2	44.6	57.9
JPEG 90	99.9	99.9	75.9	79.9	65.70	35	58.5
JPEG 80	99.9	99.9	49.3	46.9	24.4	12.1	29.6
JPEG 70	100	99.8	19	25.2	6	3.7	10.3

Se realizó una comparativa en términos de la tasa de verdaderos positivos y la tasa de verdaderos negativos, debido a que, Wang et al., [17] presenta así sus resultados. En la Tabla 6.10 se observa la comparación realizada, obteniendo mejores resultados respecto a imágenes con ruido gaussiano, y logrando mejores resultados, en general, en el conjunto de imágenes generadas con el modelo de ProGAN.

Tabla 6.10. Comparación en términos de TPR/TNR con el estado del arte.

Operación	ProGAN		StyleGAN2		StyleGAN3	
	Modelo propuesto	Wang et al. [17]	Modelo propuesto	Wang et al. [17]	Modelo propuesto	Wang et al. [17]
JPEG90	79.9/100	69.1/100	65.7/100	100/100	58.5/100	69/100
JPEG80	46.9/100	55.8/100	24.4/100	100/100	29.5/100	61.1/100
JPEG70	25.2/100	47.4/100	6/100	100/100	10.2/100	55.2/100
Ruido gaussiano	96.8/100	14/90	86.5/100	41.1/90	58.3/100	8/90
Redimensión 50%	78.2/100	85.4/99.2	53/100	100/99.2	24/100	71.2/99.2
Desenfoque gaussiano 3x3	99.5/100	72.2/99.3	97.8/100	100/99.3	71.7/100	73.3/99.3
Desenfoque gaussiano 5x5	98.8/100	63.1/99.3	81.8/100	100/99.3	41.4/100	65/99.3
Filtro de mediana 3x3	100/100	74.4/99.1	100/100	99/99.1	93.3/100	35.2/99.1

En la Tabla 6.11 se muestra la comparación con los trabajos de Chen et al. [71] y Xue et al. [79], los cuales utilizan la métrica de exactitud para evaluar sus resultados. Esto sugiere que el modelo desarrollado podría mejorarse, por ejemplo, incorporando muestras con imágenes en condiciones distintas a las generadas originalmente, añadiendo así mayor complejidad a los datos de entrenamiento. Además, es importante resaltar que el análisis de características globales y el análisis de color o las características residuales, resulta fundamental para tareas de clasificación más complejas, especialmente cuando ambas clases comparten una alta similitud, como se observa en los estudios con los que se realiza la comparación.

Tabla 6.11. Comparación en términos de exactitud con el estado del arte.

Operación	Modelo propuesto	Chen et al. [71]	Xue et al. [79]
JPEG90	73.5	99.5	99.6
JPEG70	37.7	96.6	96.8
Corrección gamma 1.1	71.5	99.6	99.5
Filtro de mediana 3x3	98	98.4	98.5
Desenfoque gaussiano 3x3	90.8	98.3	99.2
Desenfoque gaussiano 5x5	76.2	96.3	96.9
Ruido gaussiano	81.4	99.5	99.4

Escenario 3: Análisis de imágenes generadas mediante modelos GAN, que no contienen rostros humanos

En la revisión del estado del arte se encontraron trabajos como [64], [65], [70], [75], los cuales analizan imágenes que no necesariamente contienen rostros humanos, concentrándose en inferir si la imagen fue generada con tecnología GAN o no.

El procedimiento es utilizar el modelo obtenido para verificar su robustez frente a imágenes generadas por las mismas arquitecturas, con un modelo generador diferente. Existen ejemplos de conjuntos en los cuales generan imágenes de animales, como perros, gatos y animales salvajes, utilizando para su entrenamiento conjuntos como *Animal Faces High Quality* [100]. Además de, conjuntos como *LSUN* [99], el cual contiene imágenes de escenarios como, iglesia, recámaras y salas de estar, las cuales se han utilizado para entrenar modelos de *ProGAN*, *StyleGAN* y *StyleGAN2*.

Resultados

Se tiene un total de 9,000 imágenes, que constan de rostros de animales y escenas de dormitorios e iglesias, provenientes de 9 modelos GAN; de los cuales 6 modelos fueron entrenados con imágenes de gatos o de animales salvajes, utilizando los conjuntos *LSUN* [99] y *AFHQ* [100] en su primer y segunda versión.

Se planeó utilizar el conjunto original *LSUN* para tener una base sobre cómo el modelo se comportaría frente a imágenes naturales de escenarios diversos. Sin embargo, no se encontró el conjunto original en dimensiones superiores a los 128x128 píxeles. Realizar pruebas con imágenes con dicha dimensión fueron descartadas ya que es alrededor del 35% del tamaño de la entrada, lo que provocaría resultados bajos al contener menos información de la deseada. Respecto a las arquitecturas *StyleGAN2-ADA* y *StyleGAN3*, no se encontraron modelos generadores de imágenes de escenas o escenarios.

Los resultados se pueden observar en la Tabla 6.12. En donde se destaca la robustez respecto a las imágenes con rostros de animales alcanzando valores de 99%; por otro lado, se puede comprobar que el modelo entrenado únicamente con rostros humanos no es robusto a imágenes de escenarios varios, como lo son iglesias y dormitorios.

Tabla 6.12. Exactitud del modelo en los conjuntos probados; con imágenes que no contenían rostros.

Conjunto	Exactitud (%)
AFHQv2	97.5
ProGAN-LSUN-Cat	99.5
StyleGAN-LSUN-Cat	92.8
StyleGAN2-Cat	96.9
StyleGAN2-ADA-afhqcatt	97.6
StyleGAN3-afhqv2	99
ProGAN-LSUN-bedroom	15.9
StyleGAN-church	7.9
StyleGAN2-church	11.2

Escenario 4: Análisis de imágenes de un modelo GAN 3D

En el estudio del estado del arte no se encontraron pruebas utilizando imágenes generadas por la arquitectura *EG3D* [97], sin embargo, está presente en la revisión sobre las principales redes generadoras actuales, la cual, al generar una imagen, primero se obtiene un modelo en 3 dimensiones, sobre el cual se integran las texturas para obtener la imagen final.

Probar con este conjunto de imágenes tiene como objetivo comprobar la robustez del modelo desarrollado frente a las imágenes generadas. Para ello, se realizaron pruebas con un conjunto de 1,000 imágenes de rostros humanos generadas por *EG3D*.

Resultados

Para las pruebas realizadas, se generaron 1,000 imágenes utilizando el modelo ‘*ffhq512-128*’ [114] con la arquitectura de *EG3D* [97] el cual se encuentra disponible para su descarga en el repositorio del trabajo³. Las imágenes contienen rostros humanos, de un tamaño de 512x512 píxeles, al analizar el conjunto con el modelo se obtuvo 90.70% de exactitud en las imágenes generadas (Figura 6.1).

³ <https://github.com/NVlabs/eg3d>

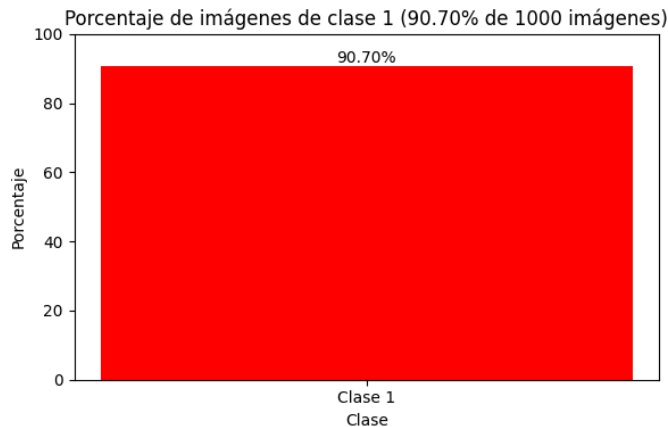


Figura 6.1. Resultados del escenario 4.

Esto indica que el modelo desarrollado generaliza bien frente a rostros generados por un modelo generador reciente. Las pruebas realizadas son valiosas para establecer un marco de referencia que pueda ser útil en investigaciones futuras sobre la robustez de las imágenes generadas por este modelo.

Escenario 5: Análisis de imágenes generadas por un modelo de difusión

El objetivo es evaluar si el modelo es robusto o no, frente a imágenes generadas por un modelo de difusión. Para realizar la experimentación, el modelo utilizado es *Stable Diffusion* [21] en su versión 1.4. Hasta donde se encontró en la revisión de la literatura, este tipo de experimentación no se encuentra presente en el estado del arte.

Resultados

El objetivo es determinar la generalización del modelo final evaluándolo con imágenes generadas con el modelo *Stable Diffusion* [21] en su versión 1.4, utilizando el estímulo textual '*a highly detailed and photorealistic portrait of a person, diverse ethnicity, various ages, ultra-realistic lighting, expressive emotions*' generadas a un tamaño de 512x512 píxeles, generando un total de 1,000 imágenes, en la Figura 6.2 se muestran dos ejemplos de las imágenes obtenidas.



Figura 6.2. Ejemplos de rostros generados con el modelo de difusión.

Para el sistema, la clase 1 corresponde a imágenes generadas mediante modelos GAN. Los resultados de esta prueba se muestran en la Figura 6.3, la cual muestra haber obtenido 29% de exactitud en la prueba realizada. Es decir, el modelo final, al no tener en el conjunto de entrenamiento información de este tipo, no es capaz de reconocer las imágenes generadas por el modelo *Stable Diffusion*, como “artificiales”.

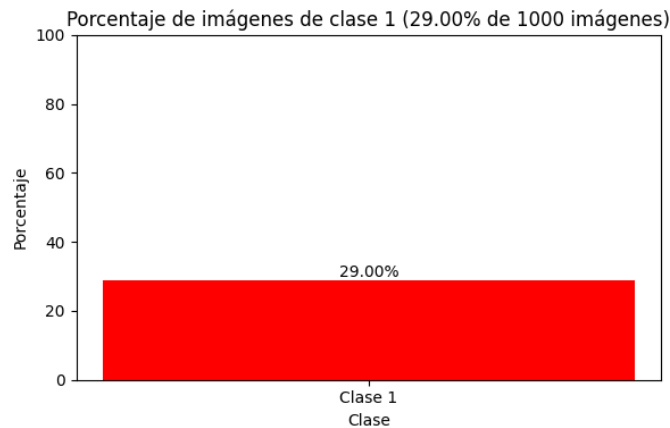


Figura 6.3. Resultados del escenario 5.

Escenario 6: Análisis de imágenes de rostros restaurados mediante un modelo GAN

El objetivo es evaluar la respuesta del modelo propuesto cuando a un rostro natural es restaurado mediante un modelo GAN enfocado específicamente en imágenes de rostros. Se espera que el modelo identifique a la imagen restaurada como “rostro natural”; ya que la restauración no debe introducir artefactos o creaciones de nuevos rasgos en la imagen.

Se utiliza el modelo *GFPGAN* [115] para restaurar 1,000 imágenes, usando imágenes del conjunto *FFHQ* habiendo reducido previamente su resolución a una menor de la original.

Resultados

En este escenario, se desea evaluar el modelo final frente a imágenes de rostros reales restaurados. Para ello, se utilizaron 1,000 imágenes del conjunto *FFHQ* como entrada. A dichas imágenes originales se les redujo la resolución a 64x64 píxeles y posteriormente fueron restauradas mediante el modelo *GFPGAN* [115]; en la Figura 6.4, se observan ejemplos de los tres tipos de imágenes: las originales, con resolución reducida y, las restauradas.



Figura 6.4. Izquierda: Imágenes originales. Centro: Imágenes a 64x64. Derecha: Imágenes restauradas.

Los resultados (ver Figura 6.5) muestran el buen desempeño del modelo final, ya que el 99.8% de las imágenes naturales restauradas mediante el modelo GAN, son percibidas como naturales. Y sólo el 0.20% de las imágenes analizadas son clasificadas erróneamente; es decir, solamente 20 imágenes fueron asignadas a la clase generada por GAN.

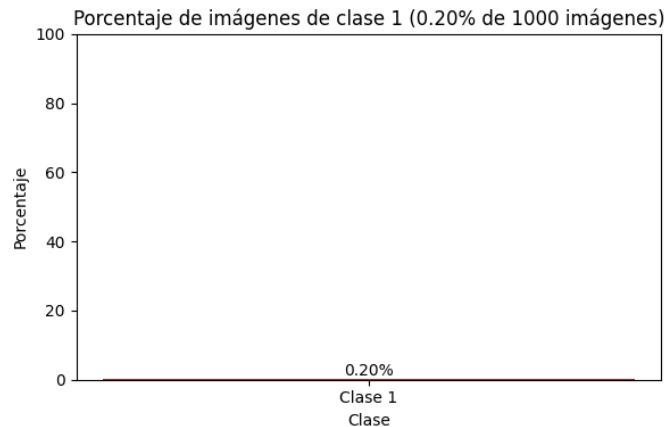


Figura 6.5. Resultados del escenario 6.

6.4. Análisis de los resultados

El objetivo principal del proyecto es desarrollar un modelo que pueda identificar si un rostro en una imagen es generado artificialmente o no, objetivo que se cumplió con éxito mediante el análisis exhaustivo del estado del arte y el desarrollo de un modelo con tecnología *transformer*, logrando una exactitud promedio del 98.39% en la clasificación. El modelo, basado en la arquitectura *MViTv2*, se entrenó con conjuntos de datos extensos y diversos, incluyendo *FFHQ*, *CelebA-HQ*, como conjuntos de datos reales, *ProGAN* y *StyleGAN2* siendo de los conjuntos generados artificialmente.

La investigación se extendió más allá de la clasificación de rostros generados por GAN en condiciones comunes, explorando la robustez del modelo final frente a diferentes condiciones y tipos de imágenes. El modelo mostró una alta precisión en la detección de imágenes generadas por modelos GAN, incluyendo *StyleGAN3*, alcanzando 94.8% de exactitud. Sin embargo, la robustez se vio afectada al aplicar técnicas de procesamiento a las imágenes, como la compresión JPEG y el desenfoque gaussiano. A pesar de esto, se obtuvieron buenos resultados en la detección de imágenes con ruido gaussiano y generadas por el modelo *ProGAN*.

Al analizar imágenes generadas por GANs que no contenían rostros humanos, el modelo mostró un buen rendimiento en la detección de imágenes con animales, especialmente al usar conjuntos de datos como *AFHQv2* y *ProGAN-LSUN-Cat*. Sin embargo, la precisión disminuyó considerablemente al evaluar imágenes generadas de escenarios como iglesias y dormitorios, lo que sugiere que el modelo está

optimizado para la detección de rostros y su rendimiento se ve afectado al analizar otro tipo de contenido.

La experimentación con las imágenes generadas por un modelo GAN, no incluido en la propuesta inicial como *EG3D*, logró resultados prometedores, con una exactitud del 90.70%. Esto sugiere que el modelo tiene la capacidad de generalizar a nuevas arquitecturas GAN.

La inclusión de técnicas de aumento de datos, como *Deep Dream* y deformación de imágenes, mejoró significativamente el rendimiento del modelo, especialmente al combinarse con técnicas tradicionales de aumento. La técnica de deformación de imágenes, en particular, mostró tener un alto impacto en la precisión, aumentando en 28 puntos porcentuales la exactitud frente a imágenes generadas por *StyleGAN3*.

La clasificación con imágenes generadas por modelos de difusión, como *Stable Diffusion*, reveló una falta de robustez del modelo debido quizá a su ausencia en el entrenamiento.

La investigación también exploró el impacto de las técnicas de restauración de rostros en la clasificación del modelo. Las imágenes naturales restauradas mediante *GFPGAN* fueron clasificadas como naturales por el modelo, lo cual era lo esperado, lo que indica que este tipo de GAN puede mantener la naturalidad del rostro.

Capítulo VII: Conclusiones

En respuesta al desafío que representa la creciente proliferación de rostros artificiales y la dificultad para distinguirlos de los reales, esta investigación culminó con el desarrollo de un modelo de detección altamente eficaz. La solución, diseñada para abordar directamente la necesidad de una herramienta precisa contra la desinformación y la manipulación de identidades, alcanzó una exactitud promedio del 98.39% en la clasificación de imágenes. Este resultado se fundamentó en la selección y optimización de la arquitectura de transformador visual MViT2, demostrando su idoneidad para esta tarea.

La construcción de este modelo se basó en una metodología rigurosa que enfrentó el problema desde múltiples ángulos. Se entrenó al sistema utilizando una combinación de conjuntos de datos de rostros reales (*FFHQ* y *CelebA-HQ*) y sintéticos (*ProGAN* y *StyleGAN2*), asegurando una base de aprendizaje diversa y representativa. El rendimiento fue potenciado mediante modificaciones clave en los parámetros de entrenamiento, como la implementación de la función de pérdida *Focal Loss* y el optimizador *AdamP*, que fueron determinantes para mejorar la capacidad del modelo de diferenciar entre las clases.

Una de las contribuciones centrales de este trabajo, y un factor clave para resolver el problema de la detección de generadores cada vez más realistas, fue la introducción de técnicas de aumento de datos no convencionales, como *Deep Dream* y la deformación de imágenes. Estas técnicas resultaron cruciales para el éxito del modelo, al simular los artefactos y texturas anómalas que caracterizan a las imágenes generadas por GANs, entrenando al sistema para reconocer estas imperfecciones sutiles. Su impacto fue notable, logrando un incremento en la precisión de hasta 28 puntos porcentuales en la clasificación de imágenes del avanzado generador *StyleGAN3*.

Al evaluar la solución, el modelo demostró una sólida capacidad de generalización, un aspecto fundamental para enfrentar la rápida evolución de las tecnologías de generación. Alcanzó una exactitud del 90.70% frente a imágenes de la arquitectura *EG3D*, la cual no formaba parte de su entrenamiento original. Sin embargo, se identificó una limitación importante en su capacidad para detectar rostros generados por modelos de difusión como *Stable Diffusion*, lo que constituye una clara vía para

trabajos futuros. Adicionalmente, se determinó que el modelo clasifica correctamente como "naturales" los rostros reales restaurados con *GFPGAN*, demostrando su habilidad para diferenciar entre una mejora de imagen y una síntesis completa desde cero.

En definitiva, este trabajo no solo cumple con el objetivo de crear una herramienta robusta y eficaz para el problema de la detección de rostros artificiales, sino que también establece que la arquitectura *MViTv2*, enriquecida con métodos de aumento de datos específicos, es una base sólida para el futuro de la detección de contenido sintético.

7.1. Objetivos logrados

El objetivo general del proyecto el cual es desarrollar un modelo que identifique si el rostro en una imagen es generado artificialmente o no, mediante el análisis de imágenes, fue cumplido al 100%. Para ello, primero se realizó un estudio de las diferentes técnicas utilizadas en el estado del arte, se evaluaron doce arquitecturas de la tecnología *transformer*. Finalmente, se seleccionó la arquitectura *MViTv2* como la de mejor desempeño, la cual se modificó y evaluó, obteniendo mejores resultados. En la Tabla 7.1, se listan los objetivos específicos junto con las actividades realizadas para su cumplimiento.

Tabla 7.1. Objetivos específicos cumplidos.

Objetivo	Detalles de su cumplimiento
Estudiar los diferentes enfoques que permiten la caracterización de un rostro natural y uno artificial, mediante el análisis del estado del arte.	Se realizó una exhaustiva revisión de la literatura, analizando aproximadamente 50 trabajos relacionados con la clasificación y detección de imágenes de rostros naturales y generados artificialmente. Se compararon las métricas reportadas, las técnicas utilizadas y los desafíos identificados en el estado del arte, lo cual se detalla en el capítulo 3.
Analizar herramientas y redes generadoras de rostros artificialmente.	Se llevó a cabo un análisis de las principales arquitecturas y herramientas generadoras de rostros artificiales, incluyendo GANs como <i>ProGAN</i> , <i>StyleGAN</i> (y sus variantes <i>StyleGAN2</i> , <i>StyleGAN2-ADA</i> , <i>StyleGAN3</i>) y modelos de difusión como <i>Stable Diffusion</i> , que sirvió como base para la selección de conjuntos de datos y la planificación de la experimentación. Esta información se aborda en los capítulos 2, 3 y 4.
Contar con conjuntos de rostros creados con herramientas de Inteligencia Artificial y rostros reales.	Se recopilaron y organizaron los conjuntos de datos necesarios, incluyendo conjuntos de rostros reales como <i>FFHQ</i> y <i>CelebA-HQ</i> , y conjuntos de rostros generados artificialmente por diversas arquitecturas GAN. Se obtuvieron 1,000 imágenes por conjunto de <i>StyleGAN2-ADA</i> y <i>StyleGAN3</i> siguiendo las instrucciones de sus repositorios oficiales, como se describe en el Capítulo 4.
Implementar una red neuronal para el reconocimiento entre rostros naturales y aquellos creados mediante herramientas de inteligencia artificial.	Se implementaron y evaluaron diversas arquitecturas de redes neuronales, centrándose en los transformadores visuales. Tras una fase de experimentación inicial con 12 modelos, se seleccionó la arquitectura <i>MViTv2</i> como la de mejor desempeño, sobre la cual se realizaron diversas modificaciones, así como la optimización de hiperparámetros claves como el <i>Dropout</i> , la función de pérdida, el optimizador y las tasas de aprendizaje, detallado en el Capítulo 5.
Evaluar el rendimiento del sistema desarrollado de manera cuantitativa a través de las métricas tradicionales en el reconocimiento de patrones.	La evaluación cuantitativa del modelo final se realizó utilizando métricas estándar en el reconocimiento de patrones como exactitud, precisión, sensibilidad (<i>TPR</i>), especificidad (<i>TNR</i>), puntuación F1 y área bajo la curva <i>ROC</i> . Los resultados obtenidos, incluyendo la evaluación en seis escenarios diferentes para medir la robustez del modelo ante diversas condiciones y tipos de imágenes, éstos se presentan y analizan en el Capítulo 5.

7.2. Contribuciones

En esta investigación se evaluaron doce arquitecturas de transformadores visuales con el objetivo de identificar la más adecuada para la clasificación de rostros naturales y generados artificialmente. Tras una serie de pruebas, se decidió trabajar con la arquitectura *MViTv2*, la cual se implementó por primera vez en el campo de la detección de imágenes generadas mediante tecnología GAN. Para optimizar su desempeño, se ajustaron parámetros clave como el *dropout*, la función de pérdida, los algoritmos de optimización y las tasas de aprendizaje automáticas, logrando un modelo con una capacidad de generalización superior.

Adicionalmente, se llevaron a cabo experimentos con nuevos conjuntos de imágenes generadas, las cuales no se habían considerado previamente en el estado del arte. Estos conjuntos incluyen imágenes producidas por *EG3D*, modelos de difusión como *Stable Diffusion* versión 1.4 y pruebas específicas para la restauración de rostros utilizando *GFPGAN*. Esta ampliación permitió evaluar el modelo en un rango más amplio de escenarios, incrementando la diversidad de datos para su entrenamiento y prueba.

En este trabajo, las principales contribuciones son el uso de técnicas de aumento de datos como Deep Dream y la deformación de imágenes, que mejoraron el rendimiento del modelo de forma notable. Los resultados mostraron que, al incluir estas técnicas durante el entrenamiento, se logró un aumento del 19 y 28 puntos porcentuales en la precisión al trabajar con el conjunto de datos StyleGAN3. Además, la incorporación de modelos de difusión y la restauración de rostros con GFPGAN abren la puerta a nuevas líneas de investigación en este campo, sentando un precedente para futuros estudios.

7.3. Trabajos futuros

Para futuras investigaciones, es importante proponer distintas líneas de trabajo que permitan fortalecer y ampliar los resultados obtenidos hasta el momento. Una de las principales áreas a considerar es la mejora de la robustez del modelo. Se sugiere investigar métodos que incrementen su capacidad para detectar imágenes generadas por modelos de difusión. Una posible estrategia sería trabajar con técnicas de procesamiento de imágenes que simulen las características propias de las imágenes generadas por este tipo de modelos. Asimismo, sería de gran utilidad incluir imágenes con post-procesamiento adicional, lo que ayudaría a evaluar el

rendimiento del modelo en condiciones no controladas y, con ello, mejorar su desempeño en escenarios más realistas.

Otra línea de trabajo relevante es ampliar el uso de técnicas de aumento de datos, sería conveniente realizar nuevos experimentos con métodos como *Deep Dream* y deformación de imágenes, variando sus parámetros y combinaciones para optimizar los resultados. También se podrían explorar otras técnicas de aumento que permitan simular artefactos y patrones visuales característicos de las imágenes generadas por GANs, mejorando así la capacidad del modelo para generalizar frente a diferentes tipos de datos.

En lo que respecta a nuevas arquitecturas y enfoques, resulta interesante evaluar otras alternativas de redes neuronales, tanto las de tipo convolucional como aquellas basadas en transformadores, con el fin de comparar su desempeño en la detección de imágenes generadas artificialmente. Además, se podría combinar la detección basada en aprendizaje profundo con el análisis de características físicas y fisiológicas de los rostros, como la simetría, la forma de la pupila o el color del iris, lo que podría ofrecer una mayor precisión en la clasificación.

Por otro lado, es importante considerar aplicaciones prácticas y realizar análisis adicionales. Un proyecto interesante sería el desarrollo de una plataforma o herramienta que permita detectar rostros generados artificialmente en contextos reales, como redes sociales. Al mismo tiempo, conviene realizar análisis interpretativos de las decisiones del modelo para identificar qué características son más influyentes en la clasificación, utilizando herramientas que permitan localizar las zonas de mayor interés en las imágenes. También sería relevante evaluar el desempeño del modelo en la detección de rostros generados parcialmente, es decir, aquellos donde solo una pequeña parte de la imagen es artificial.

Por último, se recomienda ampliar el uso de conjuntos de datos generados con diferentes técnicas y arquitecturas de redes generativas. Esto contribuiría a mejorar la capacidad de generalización del modelo y a hacerlo más robusto frente a una mayor variedad de escenarios.

7.4. Actividades académicas

Se publicó el artículo “*Eye Feature Segmentation for the Identification of Artificially Generated Faces*”, además de la presentación del mismo en el *Congreso Mexicano de Inteligencia Artificial 2025*. Un fragmento del trabajo, así como el reconocimiento obtenido se muestran en la Figura 7.1.

**Eye Feature Segmentation for the Identification
of Artificially Generated Faces**

Martin Ulises Millán Guerrero¹, Andrea Magadán Salazar¹, and Daniela
Alejandra Moctezuma Ochoa²

¹ Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), Morelos,
México
{m23ce070 & andrea.ms}@cenidet.tecnm.mx

² Centro de Investigación en Ciencias de Información Geoespacial, Aguascalientes,
México
dmoctezuma@centrogeo.edu.mx

Abstract. The advancement of generative models has enabled the creation of synthetic faces with a high degree of realism, posing challenges in verifying their authenticity. This study proposes an approach based on the analysis of eye features to detect artificially generated face images. To achieve this, three key aspects are examined: pupil shape, iris color, and the similarity of corneal specular reflections. Additionally, a solution is proposed to improve the segmentation of the iris by connecting edges in masks with holes using the YoloV8 network, addressing a common challenge in semantic segmentation. The results show that the proposed strategy performs well, achieving an accuracy exceeding 80% in identifying artificially generated faces.



La Sociedad Mexicana de Inteligencia Artificial (SMIA)
y la Universidad Panamericana

OTORGAN ESTE CERTIFICADO A

Martin Ulises Millán Guerrero, Andrea Magadán Salazar, Daniela Alejandra Moctezuma Ochoa

por la presentación del artículo titulado:

Eye Feature Segmentation for the Identification of Artificially Generated Faces

en el XVII Congreso Mexicano de Inteligencia Artificial - COMIA 2025
Ciudad de México, México, del 12 al 16 de mayo de 2025.






Figura 7.1. Evidencia y certificado de participación.

Se realizó una estancia de investigación de 3 meses de duración en la *Universidade da Coruña*, bajo la supervisión del Dr. José Joaquim De Moura Ramos, con el tema de investigación “*Analysis of chest X-ray images for the detection of pathologies, using Transformers technology and researching data augmentation methods*”.



A quien corresponda:

Por la presente, hago constar que Martin Ulises Millan Guerrero, con pasaporte/ ID número N15064625, Investigador/a de la TECNM/Centro Nacional de investigación y desarrollo tecnológico (México), ha realizado una estancia como Investigador Visitante Pre o Postdoctoral en la Universidad de A Coruña, bajo la supervisión de José Joaquim De Moura Ramos, entre el 2025-02-10 y el 2025-05-05.

Tema de la investigación: Analysis of chest X-ray images for the detection of pathologies, using transformers technology, and researching data augmentation methods.

Y para que conste a los efectos oportunos, firmo el presente en A Coruña

Figura 7.2. Certificado de finalización de estancia.

Se participó en la “*Escuela de Inteligencia Computacional y Robótica 2023*” en la *Universidad Tecnológica Emiliano Zapata*, Morelos. Donde se impartió la ponencia “*Sistema de reconocimiento de rostros naturales y creados artificialmente*”, el reconocimiento obtenido se observa en la Figura 7.3.



Figura 7.3. Certificado de exposición.

Se asistió al “*Coloquio de Inteligencia Artificial 2024*” en Chilpancingo, Guerrero. Donde se impartió la ponencia “*Reconocimiento de rostros naturales y creados artificialmente*”, el reconocimiento obtenido se observa en la Figura 7.4.

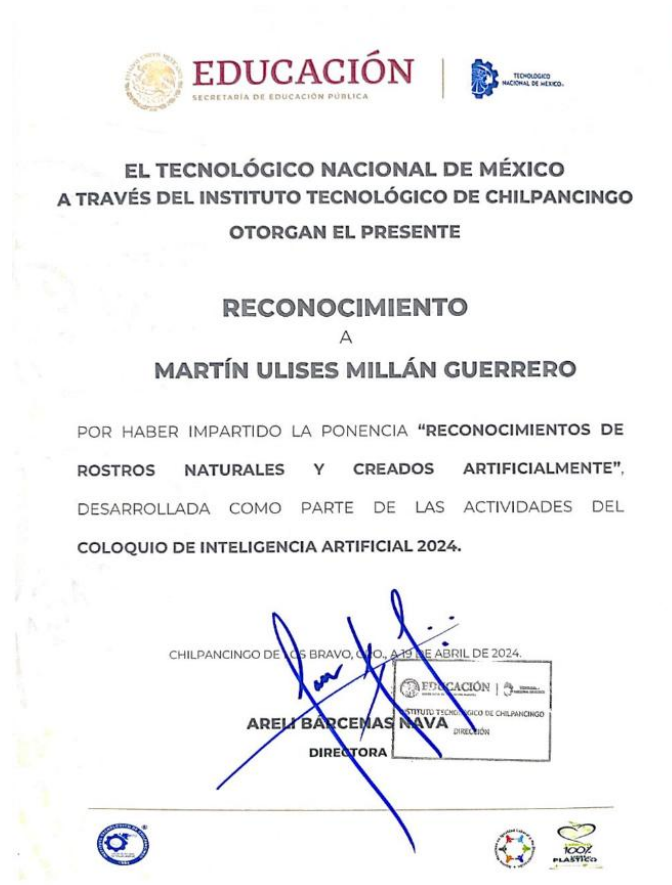


Figura 7.4. Reconocimiento obtenido en ponencia.

Se tomó el curso “YOLOv8 Performance Improvement Masterclass” en la plataforma Udeemy, impartido por el Dr. Hidayatullah, con una duración de 5.5 horas, cubriendo temas como el ajuste fino de parámetros y el aumento de datos. El certificado obtenido se muestra en la Figura 7.5.

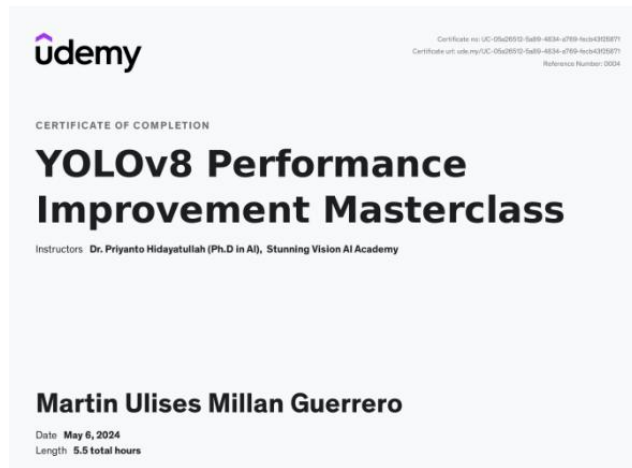


Figura 7.5. Certificado obtenido.

Se tomó el curso “PyTorch for Deep Learning BootCamp” en la plataforma Udeemy, con una duración de 52 horas, cubriendo temas como el diseño de redes neuronales, manejo de experimentación y estudio de los principales trabajos en el ámbito del aprendizaje profundo. El certificado obtenido se muestra en la Figura 7.6.

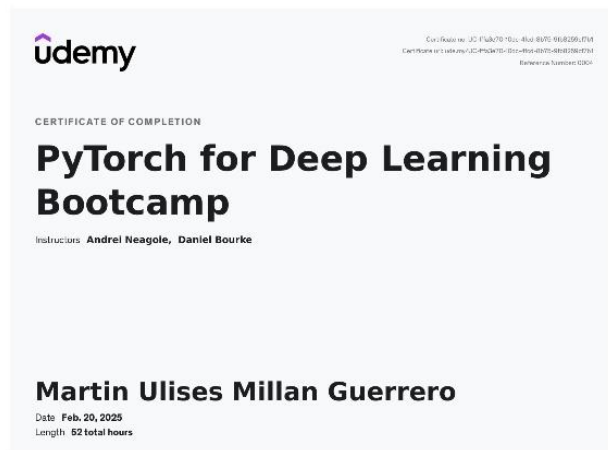


Figura 7.6. Certificado obtenido.

Referencias

- [1] S. Bond, "That smiling LinkedIn profile face might be a computer-generated fake," NPR. [Online]. Available: <https://www.npr.org/2022/03/27/1088140809/fake-linkedin-profiles>
- [2] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," Dec. 2019, [Online]. Available: <http://arxiv.org/abs/1912.04958>
- [3] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," Dec. 2018, [Online]. Available: <http://arxiv.org/abs/1812.04948>
- [4] "timm." Accessed: Sep. 11, 2024. [Online]. Available: <https://huggingface.co/docs/timm/index>
- [5] "Aspectos de un Sistema de Visión Artificial," 2005. [Online]. Available: https://www.infopl.net/files/documentacion/vision_artificial/infopl_net_Aspectos_de_un_Proyecto_de_Vision_Artificial.pdf
- [6] J. Ordieres-Meré, M. C. Limas, F. Javier Martínez De Pisón Ascacibar, and F. Alba-Elías, "Técnicas y algoritmos básicos de visión artificial Recurso electrónico-En línea," 2006. [Online]. Available: <https://www.researchgate.net/publication/231521316>
- [7] R. C. González and R. E. Woods, *Tratamiento digital de imágenes (Spanish Edition)*, 1st ed., vol. 1. Ediciones Díaz de Santos, S.A., 1996.
- [8] F. Chollet, "Deep Learning with Python," 2018. [Online]. Available: <https://tanthiamhuat.files.wordpress.com/2018/03/deeplearningwithpython.pdf>
- [9] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning."
- [10] L. Perez and J. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," Dec. 2017.
- [11] M. Ruiz Marín, J. Carlos Rodríguez Uribe, and J. Carlos Olivares Morales, "Una mirada a la biometría," 2009.
- [12] I. J. Goodfellow *et al.*, "Generative Adversarial Networks," Jun. 2014, [Online]. Available: <http://www.github.com/goodfeli/adversarial>
- [13] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: state-of-the-art, open challenges,

- countermeasures, and way forward,” *Applied Intelligence*, vol. 53, no. 4, pp. 3974–4026, Feb. 2023, doi: 10.1007/s10489-022-03766-z.
- [14] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” Dec. 2013, [Online]. Available: <http://arxiv.org/abs/1312.6114>
 - [15] S. Bond, “AI-generated fake faces are being used by bad actors online: NPR,” NPR. [Online]. Available: <https://www.npr.org/2022/12/15/1143114122/ai-generated-fake-faces-have-become-a-hallmark-of-online-influence-operations>.
 - [16] C. Pasquini, F. Laiti, D. Lobba, G. Ambrosi, G. Boato, and F. De Natale, “Identifying Synthetic Faces through GAN Inversion and Biometric Traits Analysis,” *Applied Sciences (Switzerland)*, vol. 13, no. 2, Jan. 2023, doi: 10.3390/app13020816.
 - [17] J. Wang, B. Tondi, and M. Barni, “An Eyes-Based Siamese Neural Network for the Detection of GAN-Generated Face Images,” *Frontiers in Signal Processing*, vol. 2, Jul. 2022, doi: 10.3389/frsip.2022.918725.
 - [18] H. Guo, S. Hu, X. Wang, M. C. Chang, and S. Lyu, “EYES TELL ALL: IRREGULAR PUPIL SHAPES REVEAL GAN-GENERATED FACES,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 2904–2908. doi: 10.1109/ICASSP43922.2022.9746597.
 - [19] S. Hu, Y. Li, and S. Lyu, “Exposing GaN-generated faces using inconsistent corneal specular highlights,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 2500–2504. doi: 10.1109/ICASSP39728.2021.9414582.
 - [20] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” 2020. [Online]. Available: <https://github.com/hojonathanho/diffusion>.
 - [21] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” Dec. 2021.
 - [22] A. Mordvintsev, C. Olah, and M. Tyka, “Inceptionism: Going deeper into neural networks,” *Google research blog*, vol. 20, no. 14, p. 5, 2015.
 - [23] G. Wolberg, *Digital Image Warping*, 1st ed. Washington, DC, USA: IEEE Computer Society Press, 1994.
 - [24] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, 1st ed. Springer New York, 2013. doi: 10.1007/978-1-4614-6849-3.

- [25] D. Martín Basso, “Propuesta de Métricas para Proyectos de Explotación de Información”.
- [26] H. Dalianis, “Evaluation Metrics and Evaluation,” in *Clinical Text Mining*, Springer International Publishing, 2018, pp. 45–53. doi: 10.1007/978-3-319-78503-5_6.
- [27] “Curvas ROC,” MATLAB & Simulink. Accessed: May 15, 2024. [Online]. Available: <https://la.mathworks.com/discovery/roc-curve.html>
- [28] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” Oct. 2020.
- [29] Y. Li *et al.*, “MViTv2: Improved Multiscale Vision Transformers for Classification and Detection,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2022, pp. 4794–4804. doi: 10.1109/CVPR52688.2022.00476.
- [30] H. Fan *et al.*, “Multiscale Vision Transformers,” Apr. 2021.
- [31] L. J. Peniche Ruiz, “Aplicación de celdas holográficas sobre mapas de detalle para el reconocimiento de rostros,” Centro Nacional de Investigación y Desarrollo Tecnológico, 1998.
- [32] M. G. Medina Barrera, “Metodología para el Reconocimiento de Fonemas en el Idioma Español mediante Percepción Visual por Computadora,” Centro Nacional de Investigación y Desarrollo Tecnológico, 1998.
- [33] A. Magadan Salazar, “Reconocimiento de rostros invariante a expresiones faciales,” Centro Nacional de Investigación y Desarrollo Tecnológico, 1999.
- [34] A. Herrera Peralta, “SEGUIMIENTO DE OBJETOS ALFAFLEXIBLES EN SECUENCIAS DE IMÁGENES,” CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO, 2003.
- [35] D. A. Paredes Soto, “Seguimiento y Caracterización de Componentes del Rostro para la Detección de Expresiones Faciales,” Centro Nacional de Investigación y Desarrollo Tecnológico, 2009.
- [36] M. Á. Zúñiga García, “Método para el reconocimiento de rostros mediante parámetros fractales y RNA’s,” Centro Nacional de Investigación y Desarrollo Tecnológico, 2011.
- [37] R. A. Morfin Villatoro, “Caracterización de Expresiones Faciales con un Sensor RGBD,” Centro Nacional de Investigación y Desarrollo Tecnológico, 2016.

- [38] C. A. Vázquez Rodríguez, “Valoración automática de la motivación y atención del estudiante mediante expresiones faciales,” CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO, 2017.
- [39] W. Valderrama Cardenas, “Reconocimiento automático del rostro para verificación de identidad para evaluación en línea,” Centro Nacional de Investigación y Desarrollo Tecnológico, 2019.
- [40] W. Valderrama Cardenas, “Detección de suplantación de identidad en imágenes faciales,” Centro Nacional de Investigación y Desarrollo Tecnológico, 2023.
- [41] R. Y. Zumaya Lanuza, “Mejoramiento de imágenes para sistemas de videovigilancia de baja calidad,” Centro Nacional de Investigación y Desarrollo Tecnológico, 2023.
- [42] O. de J. Trujillo Quintero, “Sistema de Visión Artificial para el Reconocimiento de la Mascarilla Facial Bien Puesta,” Centro Nacional de Investigación y Desarrollo Tecnológico, 2023.
- [43] D. Calcedo Aguirre, “Identificación de patrones atípicos en personas con movimientos involuntarios,” Centro Nacional de Investigación y Desarrollo Tecnológico, 2025.
- [44] F. Matern, C. Riess, and M. Stamminger, “Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations,” in *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, IEEE, Jan. 2019, pp. 83–92. doi: 10.1109/WACVW.2019.00020.
- [45] N. Yu, L. Davis, and M. Fritz, “Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2019, pp. 7555–7565. doi: 10.1109/ICCV.2019.00765.
- [46] Y. Fu, T. Sun, X. Jiang, K. Xu, and P. He, “Robust GAN-Face Detection Based on Dual-Channel CNN Network,” in *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, IEEE, Oct. 2019, pp. 1–5. doi: 10.1109/CISP-BMEI48845.2019.8965991.
- [47] F. Marra, C. Saltori, G. Boato, and L. Verdoliva, “Incremental learning for the detection and classification of GAN-generated images,” Oct. 2019.
- [48] R. Wang *et al.*, “FakeSpotter: A Simple yet Robust Baseline for Spotting AI-Synthesized Fake Faces,” in *Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*, 2020.

- [49] H. Mansourifar and W. Shi, “One-Shot GAN Generated Fake Face Detection,” Mar. 2020.
- [50] Z. Liu, X. Qi, and P. H. S. Torr, “Global Texture Enhancement for Fake Face Detection in the Wild,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2020, pp. 8057–8066. doi: 10.1109/CVPR42600.2020.00808.
- [51] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, “CNN-Generated Images Are Surprisingly Easy to Spot... for Now,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2020, pp. 8692–8701. doi: 10.1109/CVPR42600.2020.00872.
- [52] N. Hulzebosch, S. Ibrahimi, and M. Worring, “Detecting CNN-Generated Facial Images in Real-World Scenarios,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2020, pp. 2729–2738. doi: 10.1109/CVPRW50498.2020.00329.
- [53] L. Chai, D. Bau, S.-N. Lim, and P. Isola, “What Makes Fake Images Detectable? Understanding Properties that Generalize,” 2020, pp. 103–120. doi: 10.1007/978-3-030-58574-7_7.
- [54] H. Jeon, Y. Bang, J. Kim, and S. S. Woo, “T-GD: Transferable GAN-generated Images Detection Framework,” Aug. 2020.
- [55] M. Barni, K. Kallas, E. Nowroozi, and B. Tondi, “CNN Detection of GAN-Generated Face Images based on Cross-Band Co-occurrences Analysis,” in *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, Dec. 2020, pp. 1–6. doi: 10.1109/WIFS49906.2020.9360905.
- [56] S. Agarwal, N. Girdhar, and H. Raghav, “A Novel Neural Model based Framework for Detection of GAN Generated Fake Images,” in *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, IEEE, Jan. 2021, pp. 46–51. doi: 10.1109/Confluence51648.2021.9377150.
- [57] Z. Chen and H. Yang, “Attentive Semantic Exploring for Manipulated Face Detection,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Jun. 2021, pp. 1985–1989. doi: 10.1109/ICASSP39728.2021.9414225.
- [58] M. P. Gangan, A. K, and L. V L, “Distinguishing Natural and Computer-Generated Images using Multi-Colospace fused EfficientNet,” Oct. 2021, doi: 10.1016/j.jisa.2022.103261.
- [59] E. Nowroozi, M. Conti, and Y. Mekdad, “Detecting High-Quality GAN-Generated Face Images using Neural Networks,” Mar. 2022.

- [60] M. Zhang, H. Wang, P. He, A. Malik, and H. Liu, "Exposing unseen GAN-generated image using unsupervised domain adaptation," *Knowl Based Syst*, vol. 257, p. 109905, Dec. 2022, doi: 10.1016/j.knosys.2022.109905.
- [61] J. Lu, Y. Li, J. Zhou, B. Li, and S. Lyu, "Forensics Forest: Multi-scale Hierarchical Cascade Forest for Detecting GAN-generated Faces," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, Jul. 2023, pp. 2309–2314. doi: 10.1109/ICME55011.2023.00394.
- [62] S. McCloskey and M. Albright, "Detecting GAN-Generated Imagery Using Saturation Cues," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 4584–4588. doi: 10.1109/ICIP.2019.8803661.
- [63] X. Yang, Y. Li, and H. Qi, "Exposing GAN-synthesized Faces Using Landmark Locations," 2019, doi: 10.1145/3335203.
- [64] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "Do GANs Leave Artificial Fingerprints?," in *Proceedings - 2nd International Conference on Multimedia Information Processing and Retrieval, MIPR 2019*, Institute of Electrical and Electronics Engineers Inc., Apr. 2019, pp. 506–511. doi: 10.1109/MIPR.2019.00103.
- [65] L. Nataraj *et al.*, "Detecting GAN generated Fake Images using Co-occurrence Matrices," in *IS and T International Symposium on Electronic Imaging Science and Technology*, Society for Imaging Science and Technology, Jan. 2019. doi: 10.2352/ISSN.2470-1173.2019.5.MWSF-532.
- [66] Z. Guo, G. Yang, J. Chen, and X. Sun, "Fake face detection via adaptive manipulation traces extraction network," *Computer Vision and Image Understanding*, vol. 204, Mar. 2021, doi: 10.1016/j.cviu.2021.103170.
- [67] B. Chen, X. Ju, B. Xiao, W. Ding, Y. Zheng, and V. H. C. de Albuquerque, "Locally GAN-generated face detection based on an improved Xception," *Inf Sci (N Y)*, vol. 572, pp. 16–28, Sep. 2021, doi: 10.1016/j.ins.2021.05.006.
- [68] H. Guo, S. Hu, X. Wang, M.-C. Chang, and S. Lyu, "Robust Attentive Deep Neural Network for Exposing GAN-generated Faces," Sep. 2021, [Online]. Available: <http://arxiv.org/abs/2109.02167>
- [69] Y. Wang, V. Zarghami, and S. Cui, "FAKE FACE DETECTION USING LOCAL BINARY PATTERN AND ENSEMBLE MODELING," in *Proceedings - International Conference on Image Processing, ICIP*, IEEE Computer Society, 2021, pp. 3917–3921. doi: 10.1109/ICIP42928.2021.9506460.
- [70] M. Goebel, L. Nataraj, T. Nanjundaswamy, T. M. Mohammed, S. Chandrasekaran, and B. S. Manjunath, "Detection, attribution and localization of GAN generated images," in *IS and T International Symposium on Electronic*

Imaging Science and Technology, Society for Imaging Science and Technology, 2021. doi: 10.2352/ISSN.2470-1173.2021.4.MWSF-276.

- [71] B. Chen, X. Liu, Y. Zheng, G. Zhao, and Y. Q. Shi, "A Robust GAN-Generated Face Detection Method Based on Dual-Color Spaces and an Improved Xception," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3527–3538, Jun. 2022, doi: 10.1109/TCSVT.2021.3116679.
- [72] S. Li, V. Dutta, X. He, and T. Matsumaru, "Deep Learning Based One-Class Detection System for Fake Faces Generated by GAN Network," *Sensors*, vol. 22, no. 20, Oct. 2022, doi: 10.3390/s22207767.
- [73] J. Wang *et al.*, "GAN-generated fake face detection via two-stream CNN with PRNU in the wild," *Multimed Tools Appl*, vol. 81, no. 29, pp. 42527–42545, Dec. 2022, doi: 10.1007/s11042-021-11592-7.
- [74] F. Maher Salman and S. S. Abu-Naser, "Classification of Real and Fake Human Faces Using Deep Learning," 2022. [Online]. Available: www.ijeais.org/ijaer
- [75] Y. Ju, S. Jia, L. Ke, H. Xue, K. Nagano, and S. Lyu, "FUSING GLOBAL AND LOCAL FEATURES FOR GENERALIZED AI-SYNTHESIZED IMAGE DETECTION," in *Proceedings - International Conference on Image Processing, ICIP*, IEEE Computer Society, 2022, pp. 3465–3469. doi: 10.1109/ICIP46576.2022.9897820.
- [76] C. Liu, T. Zhu, S. Shen, and W. Zhou, "Towards Robust GAN-generated Image Detection: a Multi-view Completion Representation," Jun. 2023, [Online]. Available: <http://arxiv.org/abs/2306.01364>
- [77] M. B. I. Muafy, F. Sthevanie, and K. N. Ramadhani, "Generated AI Face Detection using Xception Model," in *2023 International Conference on Data Science and Its Applications (ICoDSA)*, IEEE, Aug. 2023, pp. 209–214. doi: 10.1109/ICoDSA58501.2023.10276634.
- [78] Y. Liu, Z. Wan, X. Yin, G. Yue, A. Tan, and Z. Zheng, "Detection of GAN generated image using color gradient representation," *J Vis Commun Image Represent*, vol. 95, p. 103876, Sep. 2023, doi: 10.1016/j.jvcir.2023.103876.
- [79] Z. Xue, X. Jiang, Q. Liu, and Z. Wei, "Global–Local Facial Fusion Based GAN Generated Fake Face Detection," *Sensors*, vol. 23, no. 2, Jan. 2023, doi: 10.3390/s23020616.
- [80] Q. Wang *et al.*, "Detecting Adversarial Faces Using Only Real Face Self-Perturbations," in *Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)*, 2023. [Online]. Available: <https://github.com/cc13qq/SAPD>

- [81] X. Liu and X. Chen, "A Survey of GAN-Generated Fake Faces Detection Method Based on Deep Learning," *Journal of Information Hiding and Privacy Protection*, vol. 2, no. 2, pp. 87–94, 2020, doi: 10.32604/jihpp.2020.09839.
- [82] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva, "Are GAN Generated Images Easy to Detect? A Critical Analysis of the State-Of-The-Art," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, Jul. 2021, pp. 1–6. doi: 10.1109/ICME51207.2021.9428429.
- [83] A. Malik, M. Kuribayashi, S. M. Abdullahi, and A. N. Khan, "DeepFake Detection for Human Face Images and Videos: A Survey," *IEEE Access*, vol. 10, pp. 18757–18775, 2022, doi: 10.1109/ACCESS.2022.3151186.
- [84] X. Wang, H. Guo, S. Hu, M.-C. Chang, and S. Lyu, "GAN-Generated Faces Detection: A Survey and New Perspectives," 2023. doi: 10.3233/FAIA230558.
- [85] X. Pan, A. Tewari, T. Leimkühler, L. Liu, A. Meka, and C. Theobalt, "DragGAN AI Tool." Accessed: Oct. 07, 2024. [Online]. Available: <https://dragganaitool.com/>
- [86] X. Pan, A. Tewari, T. Leimkühler, L. Liu, A. Meka, and C. Theobalt, "Drag Your GAN: Interactive Point-based Manipulation on the Generative Image Manifold," in *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*, New York, NY, USA: ACM, Jul. 2023, pp. 1–11. doi: 10.1145/3588432.3591500.
- [87] Generated Media Inc., "Generated Photos | Unique, worry-free model photos." Accessed: Oct. 13, 2024. [Online]. Available: <https://generated.photos/>
- [88] Playform, "Playform - AI Art Generator - Playform." Accessed: Nov. 21, 2023. [Online]. Available: <https://www.playform.io/>
- [89] Bluereach.ai, "Face Studio - The Fastest Way to Create Realistic Faces Online." Accessed: Nov. 22, 2023. [Online]. Available: <https://facestudio.app/generate>
- [90] Fotor, "Generate AI Image from Text Online for Free | Fotor." Accessed: Nov. 22, 2023. [Online]. Available: <https://www.fotor.com/images/create>
- [91] "thispersondoesnotexist.com." Accessed: Nov. 21, 2023. [Online]. Available: thispersondoesnotexist.com
- [92] A. Shinde, S. Watekar, S. Shekadar, A. Memon, and S. Gharat, "Corneal Localization for Discerning Faces using Advanced Machine Learning Algorithms of GAN," in *IDCIoT 2023 - International Conference on Intelligent Data Communication Technologies and Internet of Things, Proceedings*,

- Institute of Electrical and Electronics Engineers Inc., 2023, pp. 434–438. doi: 10.1109/IDCIoT56793.2023.10052780.
- [93] T. Karras *et al.*, “Alias-Free Generative Adversarial Networks,” 2021. [Online]. Available: <https://nvlabs.github.io/stylegan3>
 - [94] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep Learning Face Attributes in the Wild,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Dec. 2015, pp. 3730–3738. doi: 10.1109/ICCV.2015.425.
 - [95] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive Growing of GANs for Improved Quality, Stability, and Variation,” Oct. 2017.
 - [96] Z. Liu, X. Qi, and P. H. S. Torr, “Global Texture Enhancement for Fake Face Detection In the Wild.”
 - [97] E. R. Chan *et al.*, “Efficient Geometry-aware 3D Generative Adversarial Networks,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2022, pp. 16102–16112. doi: 10.1109/CVPR52688.2022.01565.
 - [98] H. Guo, S. Hu, X. Wang, M.-C. Chang, and S. Lyu, “Open-Eye: An Open Platform to Study Human Performance on Identifying AI-Synthesized Faces,” in *2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, IEEE, Aug. 2022, pp. 224–227. doi: 10.1109/MIPR54900.2022.00047.
 - [99] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, “LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop,” Jun. 2015.
 - [100] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, “StarGAN v2: Diverse Image Synthesis for Multiple Domains,” Dec. 2019.
 - [101] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, “Training Generative Adversarial Networks with Limited Data,” Jun. 2020.
 - [102] C.-F. Chen, Q. Fan, and R. Panda, “CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification,” Mar. 2021.
 - [103] S. d’Ascoli, H. Touvron, M. Leavitt, A. Morcos, G. Biroli, and L. Sagun, “ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases,” Mar. 2021, doi: 10.1088/1742-5468/ac9830.
 - [104] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang, and L. Yuan, “DaViT: Dual Attention Vision Transformers,” Apr. 2022.

- [105] H. Cai, J. Li, M. Hu, C. Gan, and S. Han, "EfficientViT: Multi-Scale Linear Attention for High-Resolution Dense Prediction," May 2022.
- [106] P. K. A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan, "FastViT: A Fast Hybrid Vision Transformer using Structural Reparameterization," Mar. 2023.
- [107] A. Hatamizadeh, H. Yin, G. Heinrich, J. Kautz, and P. Molchanov, "Global Context Vision Transformers," Jun. 2022.
- [108] S. Mehta and M. Rastegari, "Separable Self-attention for Mobile Vision Transformers," Jun. 2022.
- [109] J. Li *et al.*, "Next-ViT: Next Generation Vision Transformer for Efficient Deployment in Realistic Industrial Scenarios," Jul. 2022.
- [110] J. Chen, Q. Yu, X. Shen, A. Yuille, and L.-C. Chen, "ViTamin: Designing Scalable Vision Models in the Vision-Language Era," Apr. 2024.
- [111] "Optimizers | timmdocs." Accessed: Oct. 22, 2024. [Online]. Available: <https://timm.fast.ai/Optimizers>
- [112] NVlabs, "GitHub - NVlabs/stylegan3: Official PyTorch implementation of StyleGAN3." Accessed: Nov. 26, 2024. [Online]. Available: <https://github.com/NVLabs/stylegan3?tab=readme-ov-file>
- [113] NVlabs, "GitHub - NVlabs/stylegan2-ada-pytorch: StyleGAN2-ADA - Official PyTorch implementation." Accessed: Nov. 26, 2024. [Online]. Available: <https://github.com/NVLabs/stylegan2-ada-pytorch?tab=readme-ov-file>
- [114] NVlabs, "eg3d/docs/models.md at main · NVlabs/eg3d · GitHub." Accessed: Nov. 26, 2024. [Online]. Available: <https://github.com/NVLabs/eg3d/blob/main/docs/models.md>
- [115] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards Real-World Blind Face Restoration with Generative Facial Prior," Jan. 2021.