.

The Generalization Gap in Synthetic Image Detection: A Comparative Study of Pretrained Representations

Julieta Goria[1] y Pablo Negri[1,2]

[1] Departamento de Computación, FCEN-UBA
[2] Instituto de Investigación en Ciencias de la Computación (ICC), UBA-CONICET
jgoria@dc.uba.ar, pnegri@dc.uba.ar.

Abstract. The rapid advance of generative models, whether Generative Adversarial Networks (GANs) or diffusion models, has led to the creation of synthetic images visually indistinguishable from real ones, posing critical challenges for the security of digital information. While current detection systems achieve very high accuracy in intra-dataset evaluations, generalization to unseen synthesis mechanisms remains one of the main obstacles for their deployment. Moreover, the benchmarks commonly used to compare detectors no longer reflect the latest generators, so it remains unclear which pretrained representations actually generalize to current state-of-the-art generators. In this work, we systematically compare the generalization ability of several pretrained architectures, evaluating them on the OpenFake dataset (Livernoche et al., 2025), which includes images from state-of-the-art models such as GPT Image 1, Imagen 4 and Flux 1.1 Pro, and AIGIBench (Li et al., 2025), including images from ProGAN, R3GAN, StyleGAN-XL and SD-v1.4. The comparison includes signal-based methods (Fourier, NPR), convolutional models (VGG16, ResNet50, EfficientNet V2, MobileNet V3) and Vision Transformers (CLIP-ViT, DINOv2, Swin V2), all evaluated under a uniform protocol of frozen features with a linear classifier. Results show that Transformer-based representations generalize comparatively better to diffusion models than the other families, although their performance still degrades substantially in the cross-dataset setting.

Palabras clave. synthetic image detection, cross-dataset generalization, deepfakes, pretrained representations, Vision Transformers, generative models

La brecha de generalización en la detección de imágenes sintéticas: un estudio comparativo de representaciones preentrenadas

Abstract. El rápido avance de los modelos generativos, sea Redes Generativas Antagónicas (GANs) o modelos de difusión, ha llevado a la creación de imágenes sintéticas visualmente indistinguibles de las reales, presentando desafíos críticos para la seguridad de la información digital. Si bien los sistemas de detección actuales logran una precisión altísima en evaluaciones intra-dataset, el problema de la generalización ante nuevos mecanismos de síntesis sigue siendo uno de los principales obstáculos para su despliegue. Más aún, los benchmarks habitualmente utilizados para comparar detectores ya no reflejan a los generadores más recientes, por lo que no queda claro qué representaciones preentrenadas generalizan realmente a los generadores de última generación. En este trabajo comparamos sistemáticamente la capacidad de generalización de diversas arquitecturas preentrenadas, evaluándolas sobre el dataset OpenFake (Livernoche et al., 2025), el cual incorpora imágenes de modelos de última generación como GPT Image 1, Image 4 y Flux 1.1 Pro, y AIGIBench (Li et al., 2025), que incluye imágenes de ProGAN, R3GAN, StyleGAN-XL y SD-v1.4. La comparativa incluye métodos basados en señales (Fourier, NPR), modelos convolucionales (VGG16, ResNet50, EfficientNet V2, MobileNet V3) y Vision Transformers (CLIP-ViT, DINOv2, Swin V2), todos evaluados bajo un protocolo uniforme de features congeladas con clasificador lineal. Los resultados muestran que las arquitecturas basadas en Transformers generalizan comparativamente mejor a los modelos de difusión que las demás familias, aunque su rendimiento aún se degrada de manera considerable en el escenario cross-dataset.

Palabras clave. detección de imágenes sintéticas, generalización cross-dataset, deepfakes, representaciones preentrenadas, Vision Transformers, modelos generativos

1 Introducción

Los avances en los modelos generativos de imágenes en el último tiempo han introducido nuevos desafíos en el campo de la visión por computadora. Estos desafíos no se limitan únicamente al diseño de arquitecturas cada vez más sofisticadas o a la generación de imágenes con mayor realismo, sino que también incorporan una nueva dimensión; la capacidad de estos modelos de distinguir imágenes reales de otras generadas artificialmente (Nightingale and Farid, 2022). En este contexto, el poder de generación de estos modelos debe estar acompañado de una capacidad de discriminación comparable, para asegurar el uso responsable y ético de la tecnología (Croitoru et al., 2023; Resnik et al., 2025). Por lo tanto, el desarrollo de sistemas automatizados para la detección de estas imágenes falsas es de vital importancia. Sin embargo, la creación de los mismos presenta diversas problemáticas.

Para entrenar detectores efectivos tradicionalmente se ha recurrido a grandes cantidades de imágenes generadas por una variedad de mecanismos, representativas de la diversidad de la población (Resnik et al., 2025). Sin embargo, generar

datasets representativos a esa escala es costoso en tiempo y recursos computacionales, y depende de la disponibilidad de los generadores más avanzados. Una alternativa cada vez más extendida es apoyarse en las representaciones de modelos preentrenados, que reducen drásticamente la cantidad de imágenes de entrenamiento necesarias (Cozzolino et al., 2024); el costo se traslada entonces a la inferencia de estos backbones de gran tamaño, lo que genera preocupación por la escalabilidad y el impacto ambiental de estos métodos (Doloriel et al., 2026). Por otro lado, esta escasez de datos diversos y balanceados entre clases suele derivar en modelos con una capacidad limitada para capturar la variabilidad real de los deepfakes modernos, es decir, las imágenes o videos sintéticos generados mediante técnicas de aprendizaje profundo (Ghosh et al., 2025).

Otro problema central es la generalización. Detectar imágenes dentro de un único dominio es comparativamente sencillo, pero generalizar a otros generadores o datasets es mucho más difícil (Nadimpalli and Rattani, 2022; Xu et al., 2024): Zhu et al. (2023) muestran que la accuracy cae de 98,5% a 54-70% al entrenar y evaluar sobre generadores distintos. La dificultad se agrava por la enorme diversidad de mecanismos de generación (arquitecturas y datos de entrenamiento) y por su evolución constante, que obliga a los detectores a adaptarse a técnicas nuevas para seguir siendo efectivos.

En este trabajo se presenta un benchmark comparativo de varios modelos de detección de imágenes falsas, evaluando su eficacia y capacidad de generalización en diferentes datasets de referencia y escenarios de entrenamiento.

2 Trabajos relacionados y métodos existentes

2.1 Detectores basados en features preentrenadas

Uno de los primeros trabajos en establecer un marco sistemático para la detección de imágenes falsas fue el de Wang et al. (2020), quienes demostraron que un clasificador basado en ResNet-50, entrenado únicamente con imágenes generadas por ProGAN y aumentadas con compresión JPEG y desenfoque gaussiano, era capaz de generalizar a otros generadores basados en GANs con una precisión superior al 90%. Este resultado estableció el protocolo de evaluación que adoptaría gran parte de la literatura posterior: entrenar con un único generador y medir la capacidad de transferencia a generadores no vistos durante el entrenamiento. Sin embargo, trabajos posteriores evidenciaron que este enfoque presenta limitaciones significativas frente a los modelos de difusión, cuyas imágenes no comparten los mismos artefactos estructurales que las generadas por GANs (Cozzolino et al., 2024; Ojha et al., 2024).

Frente a estas limitaciones, Ojha et al. (2024) propusieron un cambio de paradigma. En lugar de entrenar un clasificador end to end, extrajeron representaciones de un modelo CLIP ViT-L/14, originalmente preentrenado para alinear imágenes con descripciones textuales, y las combinaron con un clasificador lineal entrenado exclusivamente sobre imágenes de ProGAN. El backbone de CLIP no recibió ningún finetuning para la tarea de detección. La idea detrás de este

enfoque es que un modelo cuyo objetivo original no guarda relación con la detección aprende representaciones más generalizables que un clasificador binario entrenado desde cero, el cual tiende a sobreajustarse a los artefactos específicos del generador con el que fue entrenado. Los resultados respaldaron esta idea: el modelo alcanzó 99,31 mAP (mean average precision) sobre GANs y 95,00 mAP sobre generadores no vistos (incluyendo difusión y autoregresivos), mejorando en +15 mAP al estado del arte previo. Este trabajo posicionó a las features de CLIP como la representación de referencia para la detección generalizable de imágenes sintéticas.

Cozzolino et al. (2024) refinaron este enfoque evaluando sistemáticamente distintas configuraciones de extracción de características y clasificación sobre las representaciones de CLIP. En particular, demostraron que las features extraídas de la penúltima capa del codificador visual de CLIP ViT-L/14, combinadas con un clasificador SVM lineal, constituyen una configuración notablemente robusta: entrenando con tan solo 1000 imágenes provenientes de un único generador, este esquema igualó al estado del arte en evaluaciones intra-distribución y lo superó en +6% AUC sobre generadores no vistos y en +3% AUC sobre imágenes post-procesadas con compresión y redimensionamiento. Estos resultados fueron validados sobre 18 modelos generativos, incluyendo DALL·E 3, Midjourney v5 y Adobe Firefly.

2.2 Otros métodos de detección

Una línea de trabajo complementaria ha explorado la detección de artefactos directamente en el dominio de la señal. Nuestro benchmark incluye un análisis basado en la transformada de Fourier, motivado por la observación de Frank et al. (2020) de que las arquitecturas generativas introducen artefactos espectrales característicos en altas frecuencias.

Si bien este enfoque ha demostrado ser efectivo para imágenes generadas por GANs, los modelos de difusión más recientes presentan una alineación espectral considerablemente mejor con las imágenes reales, lo que reduce la utilidad de las representaciones puramente frecuenciales (Doloriel et al., 2026; C. Tan et al., 2024). En esta misma línea, C. Tan et al. (2023) propusieron explotar las relaciones entre píxeles vecinos (NPR) generadas por las operaciones de up-sampling comunes tanto a GANs como a modelos de difusión. Entrenado únicamente con imágenes de ProGAN, este método alcanzó un 93,3% de accuracy promedio sobre 28 generadores, superando a UnivFD por +13,2%. Esto indica que los artefactos de bajo nivel introducidos por las operaciones de up-sampling persisten a través de generadores de distintos paradigmas y conservan poder discriminativo cuando la representación está específicamente diseñada para capturarlos.

2.3 Datasets y benchmarks de referencia

Los resultados que se pueden obtener en detección de imágenes sintéticas dependen fuertemente de los datasets utilizados, tanto para el entrenamiento como

para la evaluación. ForenSynths (Wang et al., 2020) constituyó el primer benchmark ampliamente adoptado, incluyendo imágenes de 11 generadores basados en GANs como ProGAN, StyleGAN, BigGAN y CycleGAN, pero su cobertura se limita a esta familia de arquitecturas. GenImage (Zhu et al., 2023) amplió significativamente la escala y diversidad del problema, con 2,68 millones de imágenes provenientes de 8 generadores que incluyen modelos de difusión como Stable Diffusion y Midjourney, aunque ninguno de sus generadores es posterior a 2023. Más recientemente, DF40 (Yan et al., 2024) propuso el benchmark más exhaustivo para deepfakes faciales, con 40 técnicas de generación distintas, y evidenció que datasets ampliamente utilizados como FaceForensics++ resultan insuficientes para entrenar detectores capaces de generalizar a las técnicas actuales.

Un problema transversal a estos benchmarks es su disponibilidad práctica. Como documentan Livernoche et al. (2025), la mayoría de los benchmarks existentes (incluyendo ForenSynths, GenImage, DRCT, FaceForensics++ y Celeb-DF) presentan acceso difícil o restringido, con enlaces de descarga expirados, servidores con limitaciones de velocidad o formularios de solicitud sin respuesta garantizada. Esto dificulta la reproducibilidad y limita la posibilidad de hacer comparaciones rigurosas.

3 Metodología

Table 1. Arquitecturas evaluadas y dimensión del vector de features extraído.

Familia	**Modelo**	**Dim.**	**Preentrenamiento**
Señal	Fourier	48	—
Señal	NPR	222	—
CNN	VGG16	512	ImageNet
CNN	ResNet50	2048	ImageNet
CNN	EfficientNet V2-S	1280	ImageNet
CNN	MobileNet V3 Large	960	ImageNet
Transformer	CLIP ViT-B/16	512	imagen-texto (400M pares)
Transformer	DINOv2 ViT-L/14	1024	auto-supervisado (142M imgs)
Transformer	Swin V2-S	768	ImageNet

Se seleccionaron nueve métodos de extracción de features organizados en tres familias: métodos basados en señales (Fourier y NPR), redes convolucionales preentrenadas (VGG16, ResNet50, EfficientNet V2 y MobileNet V3) y Vision Transformers (CLIP ViT, DINOv2 y Swin V2). En todos los casos se usaron las features extraídas sin fine-tuning y se entrenó un clasificador lineal, lo que permite comparar las representaciones en igualdad de condiciones. Esta decisión reduce el riesgo de sobreajuste y permite evaluar qué tan buena es cada representación para el problema de detección por sí misma, siguiendo el protocolo de Razavian et al. (2014), adoptado para detección de imágenes sintéticas por Ojha et al. (2024) y Cozzolino et al. (2024). Un resumen de las dimensiones de las features extraídas por cada modelo y cualquier preentrenamiento se puede ver en la Tabla 1.

3.1 Métodos clásicos o basados en señales

Los primeros trabajos sobre detección de imágenes falsas se basaban en la hipótesis de que las operaciones matemáticas de la síntesis de imágenes, específicamente el up-sampling u operaciones de convolución, dejaban trazas distintivas en el dominio de la frecuencia (Frank et al., 2020). Por lo tanto si bien los modelos generativos pueden lograr hiperrealismo a nivel píxeles, su habilidad para replicar las características espectrales de imágenes reales está limitada por la propia estructura de las arquitecturas neuronales (Durall et al., 2020).

Análisis basado en Fourier La detección basada en Fourier transforma las imágenes del dominio espacial $f(x, y)$ al dominio de la frecuencia $F(u, v)$, típicamente mediante la Transformada de Fourier rápida (FFT) o la transformada de coseno discreta (DCT). Frank et al. (2020) demostraron que las imágenes generadas por GANs presentan artefactos periódicos en el espectro de altas frecuencias (patrones de "cimas" o "grillas") que no aparecen en imágenes reales. Sin embargo, los modelos de difusión más recientes tienen una mejor alineación espectral con imágenes reales, lo que hace que la detección puramente frecuencial sea menos efectiva contra ellos (C. Tan et al., 2024).

Neighboring Pixel Relationships (NPR) NPR (C. Tan et al., 2023) parte de una observación concreta: las operaciones de up-sampling, presentes tanto en GANs como en modelos de difusión, introducen relaciones predecibles entre píxeles vecinos que difieren de las correlaciones locales de una imagen real. Para cada píxel se computa el residuo absoluto promedio respecto a sus vecinos en las 4 direcciones cardinales, generando un mapa de residuos que amplifica estos artefactos. A partir de este mapa se extraen features estadísticos por canal (media, desviación estándar, asimetría, curtosis, percentiles) junto con un histograma normalizado de los valores residuales. A diferencia del análisis de Fourier, que opera en el dominio de la frecuencia global, NPR captura artefactos locales a nivel de píxel, lo que lo hace potencialmente más robusto frente a generadores con buena alineación espectral.

3.2 Detectores basados en features de CNNs preentrenadas

Las CNNs entrenadas sobre ImageNet producen representaciones visuales que se han demostrado transferibles a tareas muy distintas de la clasificación de objetos original (Razavian et al., 2014). Se seleccionaron cuatro arquitecturas como extractores de features congelados: VGG16 (Simonyan and Zisserman, 2015), ResNet50 (He et al., 2015), EfficientNet V2-S (M. Tan and Le, 2021) y MobileNet V3 Large (Howard et al., 2019). En todos los casos se toma la salida de la última capa convolucional y se aplica Global Average Pooling (GAP), obteniendo un vector de features de dimensión fija: 512-d para VGG16, 2048-d para ResNet50, 1280-d para EfficientNet V2-S y 960-d para MobileNet V3 Large (véase Tabla 1), y se entrena un clasificador lineal sobre ese vector. Todas las

imágenes se redimensionan a 224×224 y se normalizan con la media y desviación estándar de ImageNet. No se realiza fine-tuning ni data augmentation.

La inclusión de estas arquitecturas responde a su reconocimiento en la literatura de detección: el protocolo de Wang et al. (2020) usa ResNet50, y GenImage (Zhu et al., 2023) evalúa varias de ellas en su benchmark cross-generador. Su presencia permite contrastar las representaciones convolucionales clásicas contra los Vision Transformers y los métodos de señales.

3.3 Detectores basados en Vision Transformers

CLIP ViT A diferencia de las CNNs, CLIP (Contrastive Language-Image Pretraining) (Radford et al., 2021) fue preentrenado con aprendizaje contrastivo sobre unos 400 millones de pares imagen-texto de internet, alineando representaciones visuales y textuales en un espacio de embedding compartido. El codificador visual es un Vision Transformer (ViT) (Dosovitskiy et al., 2021) que divide la imagen en parches de 16×16 píxeles y los procesa como una secuencia con capas de self-attention, lo que le permite capturar dependencias globales entre regiones de la imagen desde las primeras capas (a diferencia de las CNNs, cuyo campo receptivo crece gradualmente con la profundidad).

Utilizamos la variante ViT-B/16 con los pesos originales de OpenAI, extrayendo el embedding visual de 512 dimensiones producido por `encode_image`. La idea, respaldada por Ojha et al. (2024) y Cozzolino et al. (2024), es que un modelo entrenado para tareas generales aprende representaciones más transferibles que un clasificador binario especializado, que tiende a sobreajustarse a los artefactos del generador de entrenamiento.

DINOv2 DINOv2 (Oquab et al., 2024) es un Vision Transformer entrenado de forma auto-supervisada sobre un corpus curado de 142 millones de imágenes, sin etiquetas ni texto. A diferencia de CLIP, que alinea representaciones visuales y textuales, DINOv2 aprende representaciones puramente visuales optimizando la consistencia entre distintas vistas aumentadas de una misma imagen mediante destilación de conocimiento (teacher-student). Esto produce features que capturan la estructura visual de las imágenes sin depender de supervisión textual. Utilizamos la variante ViT-L/14 (Large), que produce un embedding de 1024 dimensiones a partir del token CLS. Su inclusión permite evaluar si las representaciones auto-supervisadas ofrecen ventajas frente a las contrastivas de CLIP para la detección de imágenes sintéticas.

Swin Transformer V2 Swin Transformer V2 (Liu et al., 2022) es una arquitectura jerárquica de Vision Transformer que procesa la imagen mediante ventanas locales desplazadas (*shifted windows*), reduciendo la complejidad de la self-attention de cuadrática a lineal respecto al tamaño de la imagen. A diferencia de ViT, que aplica atención global sobre toda la imagen desde la primera capa, Swin construye una representación multi-escala: las primeras capas capturan detalles finos en ventanas pequeñas, y las capas posteriores, tras fusiones

de parches (*patch merging*), capturan contexto a mayor escala. La versión V2 incorpora además un bias de posición logarítmico que mejora la transferibilidad entre resoluciones. Utilizamos la variante Swin-V2-S (Small), preentrenada sobre ImageNet, que produce un vector de 768 dimensiones. Esta arquitectura es relevante para nuestro benchmark porque OpenFake (Livernoche et al., 2025) usa un SwinV2 como su detector de referencia, lo que permite una comparación directa con sus resultados.

3.4 Datasets utilizados

En este trabajo utilizamos OpenFake (Livernoche et al., 2025), un dataset de aproximadamente 4 millones de imágenes que incorpora generadores de última generación ausentes en los benchmarks anteriores, incluyendo GPT Image 1, Imagen 4, Flux 1.1 Pro, SDXL e Ideogram. Para dimensionar la importancia de evaluar sobre estos generadores recientes: los propios autores del dataset reportan que un SwinV2-Small entrenado en OpenFake alcanza F1 = 0,99 en evaluación intra-distribución y F1 = 0,86 sobre imágenes de redes sociales, mientras que un modelo entrenado en GenImage obtiene apenas F1 = 0,08 en ese mismo escenario. Esto indica que los benchmarks anteriores ya no capturan bien las características de los generadores actuales. Cabe notar que los generadores de OpenFake pertenecen en su totalidad al paradigma de difusión y modelos autoregresivos de última generación, sin cobertura de GANs; por ello lo empleamos en el escenario intra-dataset, como medida de la dificultad de detectar a los generadores actuales más realistas. Además, OpenFake es una plataforma abierta y continuamente actualizada, lo que facilita la reproducibilidad y la incorporación de nuevos generadores.

Para la evaluación de generalización cross-paradigma, complementamos OpenFake con AIGIBench (Li et al., 2025). A diferencia de OpenFake, AIGIBench abarca dos paradigmas generativos: GANs (ProGAN, R3GAN, StyleGAN-XL) y difusión (SD-v1.4). Esta cobertura es la que permite medir la transferencia entre paradigmas (entrenar con uno y evaluar sobre otro), algo que OpenFake no posibilita al carecer de GANs. Ambos datasets se encuentran disponibles en HuggingFace.

Se pueden ver ejemplos de ambos datasets en las figuras 1 y 2.

3.5 Configuración experimental

Todos los métodos comparten el mismo pipeline de evaluación: se extraen las features con el modelo congelado y se entrena un clasificador LinearSVC de scikit-learn (C = 1.0, `class_weight=balanced`), con `CalibratedClassifierCV` para obtener probabilidades calibradas. El threshold de decisión se optimiza sobre el conjunto de validación maximizando el F1-score. No se realiza fine-tuning, data augmentation ni selección de hiperparámetros por modelo.

Cada dataset se utiliza en un pipeline de evaluación independiente (OpenFake para el escenario intra-dataset y AIGIBench para el cross-dataset), sin combinarse en un mismo conjunto de entrenamiento. Para OpenFake se utilizó un

subconjunto de 99.655 imágenes (40.000 para entrenamiento, 59.655 para test), balanceado por generador. Su tamaño no responde a una limitación de muestras, dado que OpenFake supera los 4 millones de imágenes, sino al volumen de descarga, ya que el dataset completo supera 1 TB. Para AIGIBench se utilizaron las particiones definidas por los autores del dataset Li et al., 2025, con 302.024 imágenes de entrenamiento y 212.802 de test.

Los experimentos se ejecutaron sobre un sistema con dos GPUs NVIDIA RTX 3080 (12 GB y 10 GB de VRAM), 62 GB de RAM y 48 cores de CPU, con CUDA 11.8.

4 Resultados

Los resultados se organizan en tres escenarios con complejidad creciente: primero la evaluación intra-dataset sobre OpenFake, luego la evaluación cross-dataset sobre AIGIBench entrenando solo con GANs (Setting 1), y finalmente el mismo test set pero entrenando con GANs y difusión (Setting 2). Esta progresión permite observar cómo se degrada el rendimiento cuando cambia la distribución entre entrenamiento y evaluación, y en qué medida diversificar el entrenamiento ayuda a mitigar esa caída.

4.1 Evaluación intra-dataset: OpenFake

La tabla 2 muestra los resultados de todos los métodos sobre la partición de test de OpenFake, entrenados y evaluados sobre la misma distribución de generadores.

Cuando entrenamiento y test comparten la misma distribución, se observa una jerarquía clara entre las tres familias. Los Vision Transformers lideran, con CLIP ViT alcanzando un AUC de 0,962 y un F1 de 0,899, seguido por Swin V2 (0,926) y DINOv2 (0,910). Esto es consistente con lo reportado por Ojha et al. (2024) y Cozzolino et al. (2024) sobre las representaciones de CLIP. Cabe destacar que Swin V2 es la arquitectura utilizada como detector de referencia por los propios autores de OpenFake (Livernoche et al., 2025).

Las CNNs quedan en un segundo nivel, con AUCs entre 0,836 y 0,904. ResNet50 lidera el grupo. Vale notar que la diferencia entre la mejor CNN (ResNet50, 0,904) y el peor Transformer (DINOv2, 0,910) es de apenas 0,006, lo que indica que en este escenario la ventaja de los Transformers es moderada.

Los métodos basados en señales muestran un rendimiento sustancialmente inferior, con AUCs de 0,712 (NPR) y 0,695 (Fourier). Este resultado es esperado dado que OpenFake incluye generadores de difusión recientes como GPT Image 1 e Imagen 4, que presentan una mejor alineación espectral con imágenes reales y por lo tanto dificultan la detección por artefactos de frecuencia Doloriel et al., 2026; C. Tan et al., 2024.

Table 2. Resultados sobre OpenFake (evaluación intra-dataset). Todos los modelos fueron entrenados y evaluados sobre el mismo dataset ($n_{\text{test}} = 59.655$). El *threshold* fue optimizado sobre el conjunto de validación.

Familia	Método	AUC	Accuracy	F1	Recall
Transformers	CLIP ViT	**0.962**	**0.896**	**0.899**	0.921
	Swin V2	0.926	0.848	0.848	0.846
	DINOv2	0.910	0.830	0.828	0.817
CNNs	ResNet50	0.904	0.824	0.824	0.825
	MobileNet V3	0.887	0.803	0.807	0.821
	VGG16	0.840	0.756	0.762	0.781
	EfficientNet V2	0.836	0.757	0.757	0.759
Señales	NPR	0.712	0.661	0.642	0.609
	Fourier	0.695	0.639	0.636	0.630

4.2 Evaluación cross-dataset: AIGIBench

Para evaluar la capacidad de generalización, se entrenaron todos los métodos sobre AIGIBench (Li et al., 2025) en dos configuraciones: el Setting 1 entrena solo con imágenes de GANs, mientras que el Setting 2 incorpora también imágenes de modelos de difusión. Ambos comparten el mismo test set ($n_{\text{test}} = 212.802$), que incluye imágenes tanto de GANs como de difusión, lo que permite aislar el efecto de la composición del entrenamiento. Los resultados se pueden ver en la tabla 3..

Table 3. Resultados sobre AIGIBench en ambos settings ($n_{\text{test}} = 212.802$). Setting 1: entrenamiento solo con GANs. Setting 2: entrenamiento con GANs + difusión. ΔAUC indica la diferencia respecto a OpenFake.

		Setting 1 (solo GANs)			**Setting 2 (GANs + difusión)**		
Familia	**Método**	**AUC**	**F1**	**ΔAUC**	**AUC**	**F1**	**ΔAUC**
Transformers	CLIP ViT	0.606	0.471	−0.356	**0.749**	**0.689**	−0.213
	Swin V2	**0.637**	**0.606**	−0.289	0.664	0.668	−0.262
	DINOv2	0.532	0.357	−0.378	0.631	0.564	−0.279
CNNs	ResNet50	0.552	0.476	−0.352	0.584	0.529	−0.320
	EfficientNet V2	0.568	0.528	−0.268	0.594	0.598	−0.242
	MobileNet V3	0.498	0.190	−0.389	0.553	0.405	−0.334
	VGG16	0.513	0.266	−0.327	0.573	0.387	−0.267
Señales	NPR	0.524	0.361	−0.188	0.491	0.303	−0.221
	Fourier	0.465	0.137	−0.230	0.488	0.356	−0.207

Setting 1: entrenamiento con GANs El cambio de escenario produce una caída drástica en todos los métodos. La caída promedio en AUC respecto a OpenFake es de 0,309 puntos, y varios modelos quedan cerca del azar (AUC $\approx$ 0,50). Estos números son consistentes con lo que reportan Zhu et al. (2023), donde la accuracy cae de 98,5% al rango 54-70% al evaluar sobre generadores distintos a los de entrenamiento. Dentro de esta caída generalizada, los Transformers se degradan menos que el resto. Swin V2 obtiene el mejor resultado (AUC 0,637,

F1 0,606), superando a CLIP ViT (0,606, F1 0,471). Esta inversión respecto a OpenFake es interesante: CLIP domina cuando la distribución de test coincide con la de entrenamiento, pero Swin resiste mejor cuando no. Una posible explicación es que la atención jerárquica de Swin, que procesa la imagen a múltiples escalas, captura artefactos de bajo nivel que son compartidos entre GANs y difusión, mientras que CLIP se apoya más en features semánticas de alto nivel que no transfieren tan bien en este escenario. DINOv2 sufre la mayor caída entre los Transformers (ΔAUC $= -0,378$), lo que muestra que sus representaciones auto-supervisadas no generalizan tan bien a paradigmas no vistos.

Las CNNs caen aún más: MobileNet V3 (AUC 0,498) y VGG16 (0,513) quedan esencialmente en el azar. EfficientNet V2 es la que mejor resiste (ΔAUC $= -0,268$).

En cuanto a los métodos de señales, Fourier cae a un AUC de 0,465, es decir por debajo del azar. Esto indica que los patrones espectrales aprendidos de GANs no solo no sirven para detectar imágenes de difusión, sino que resultan contraproducentes, posiblemente porque las características espectrales de los modelos de difusión son opuestas a las de las GANs (Frank et al., 2020; C. Tan et al., 2024). NPR se mantiene levemente por encima (0,524), lo que es consistente con su diseño orientado a capturar artefactos de up-sampling compartidos entre ambos paradigmas (C. Tan et al., 2023).

Setting 2: entrenamiento con GANs y difusión Agregar imágenes de difusión al entrenamiento produce una recuperación parcial del rendimiento. El caso más notable es CLIP ViT, que sube de 0,606 a 0,749 en AUC (+0,143), recuperando su posición como mejor método. Esto es consistente con los resultados de Ojha et al. (2024): las representaciones de CLIP parecen aprovechar mejor la diversidad en los datos de entrenamiento.

Swin V2 mejora menos (+0,027), lo que refuerza la idea de que su buen rendimiento en el Setting 1 se debe más a una robustez de base que a la capacidad de beneficiarse de más datos. DINOv2 sube 0,099, alcanzando un AUC de 0,631.

Las CNNs mejoran entre 3 y 6 puntos de AUC pero ninguna supera 0,594. Los métodos de señales prácticamente no se benefician de la diversificación: NPR incluso baja levemente (de 0,524 a 0,491), lo que tiene sentido si consideramos que estos métodos capturan artefactos específicos de cada familia de generadores, y al mezclar señales de GANs y difusión en el entrenamiento, las pistas pueden ser contradictorias entre sí.

5 Conclusiones

Este trabajo presentó un benchmark comparativo de nueve métodos de detección de imágenes sintéticas, evaluados con un protocolo uniforme de features congeladas y clasificador lineal sobre OpenFake y AIGIBench.

Los resultados muestran que, con los generadores disponibles actualmente, la detección intra-dataset alcanza un rendimiento muy alto, con AUCs por encima

Fig. 1. Ejemplos de clasificación sobre imágenes de OpenFake (evaluación intra-dataset). Debajo se indica el resultado de cada método (✓ = acierto, × = error).

Fig. 2. Ejemplos de clasificación sobre imágenes de AIGIBench (evaluación cross-dataset, Setting 2).

de 0,90 para la mayoría de los métodos basados en redes neuronales. Sin embargo, la generalización cross-dataset sigue siendo un problema abierto: CLIP ViT pierde más de un tercio de su rendimiento al cambiar de distribución, y la mayoría de las CNNs y métodos de señales caen al nivel del azar.

Los Vision Transformers, y en particular CLIP ViT, son la familia de representaciones más prometedora para la detección generalizable, gracias a la transferibilidad de sus features preentrenadas. Sin embargo, incluso estas representaciones no alcanzan sin datos de entrenamiento que cubran la diversidad de paradigmas generativos.

Como trabajo futuro, sería interesante explorar estrategias de adaptación de dominio que reduzcan la dependencia en datos de entrenamiento diversos, así como evaluar sobre generadores más recientes a medida que estos se incorporen a plataformas como OpenFake. También podría ser útil combinar representa-

ciones de distintas familias (por ejemplo, señales + Transformers), que podrían complementarse para mejorar la robustez frente a generadores no vistos.

Agradecimientos. Este trabajo fue realizado bajo la beca BIICC del Departamento de Computacion - Facultad de Ciencias Exactas y Naturales - Universidad de Buenos Aires.

Declaración de intereses. Durante la preparación de este trabajo, el/los autor(es) utilizaron claude.ai con el fin de mejorar el lenguaje y la legibilidad del manuscrito. Tras utilizar esta herramienta/servicio, el/los autor(es) revisaron y editaron el contenido según fuera necesario y asumen plena responsabilidad por el contenido de la publicación.

Referencias

Cozzolino, D., Poggi, G., Corvi, R., Nießner, M., & Verdoliva, L. (2024). Raising the bar of ai-generated image detection with clip. https://arxiv.org/abs/2312.00195

Croitoru, F.-A., Hondru, V., Ionescu, R. T., & Shah, M. (2023). Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(9), 10850–10869. https://doi.org/10.1109/tpami.2023.3261988

Doloriel, C. T. C., Ullah, H., Liland, K. H., Machot, F. A., & Cheung, N.-M. (2026). Towards sustainable universal deepfake detection with frequency-domain masking. https://arxiv.org/abs/2512.08042

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. https://arxiv.org/abs/2010.11929

Durall, R., Keuper, M., & Keuper, J. (2020). Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7887–7896. https://api.semanticscholar.org/CorpusID:211988680

Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., & Holz, T. (2020). Leveraging frequency analysis for deep fake image recognition. https://arxiv.org/abs/2003.08685

Ghosh, T., Seth, B., Kar, S., & Naskar, R. (2025). *Diffsynface: A demographically (age, gender and ethnicity) diverse synthetic face dataset*. https://doi.org/10.21227/01es-2p20

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. https://arxiv.org/abs/1512.03385

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). Searching for mobilenetv3. https://arxiv.org/abs/1905.02244

Li, Z., Yan, J., He, Z., Zeng, K., Jiang, W., Xiong, L., & Fu, Z. (2025). Is artificial intelligence generated image detection a solved problem? https://arxiv.org/abs/2505.12335

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., & Guo, B. (2022). Swin transformer v2: Scaling up capacity and resolution. https://arxiv.org/abs/2111.09883

Livernoche, V., Arodi, A., Musulan, A., Yang, Z., Salvail, A., Caron, G. M., Godbout, J.-F., & Rabbany, R. (2025). Openfake: An open dataset and platform toward real-world deepfake detection. https://arxiv.org/abs/2509.09495

Nadimpalli, A. V., & Rattani, A. (2022). On improving cross-dataset generalization of deepfake detectors. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 91–99.

Nightingale, S. J., & Farid, H. (2022). Ai-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, *119*(8), e2120481119. https://doi.org/10.1073/pnas.2120481119

Ojha, U., Li, Y., & Lee, Y. J. (2024). Towards universal fake image detectors that generalize across generative models. https://arxiv.org/abs/2302.10174

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., ... Bojanowski, P. (2024). Dinov2: Learning robust visual features without supervision. https://arxiv.org/abs/2304.07193

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. https://arxiv.org/abs/2103.00020

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). Cnn features off-the-shelf: An astounding baseline for recognition. https://arxiv.org/abs/1403.6382

Resnik, D., Hosseini, M., Kim, J., Epiphaniou, G., & Maple, C. (2025). Genai synthetic data create ethical challenges for scientists. here's how to address them. *Proceedings of the National Academy of Sciences of the United States of America*, *122*, e2409182122. https://doi.org/10.1073/pnas.2409182122

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. https://arxiv.org/abs/1409.1556

Tan, C., Liu, H., Zhao, Y., Wei, S., Gu, G., Liu, P., & Wei, Y. (2023). Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. https://arxiv.org/abs/2312.10461

Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., & Wei, Y. (2024). Frequency-aware deepfake detection: Improving generalizability through frequency space learning. https://arxiv.org/abs/2403.07240

Tan, M., & Le, Q. V. (2021). Efficientnetv2: Smaller models and faster training. https://arxiv.org/abs/2104.00298

Wang, S.-Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). Cnn-generated images are surprisingly easy to spot... for now. https://arxiv.org/abs/1912.11035

Xu, Y., Terhörst, P., Pedersen, M., & Raja, K. (2024). Analyzing fairness in deepfake detection with massively annotated databases. *IEEE Transactions on Technology and Society*, *5*(1), 93–106.

Yan, Z., Yao, T., Chen, S., Zhao, Y., Fu, X., Zhu, J., Luo, D., Wang, C., Ding, S., Wu, Y., & Yuan, L. (2024). Df40: Toward next-generation deepfake detection. https://arxiv.org/abs/2406.13495

Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., & Wang, Y. (2023). Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in neural information processing systems*, *36*, 77771–77782.