Generación de imágenes de ropa en Fashion-MNIST: estudio comparativo de tres configuraciones de DCGAN

- Autor: Evie Nataly Díaz Quevedo
- Afiliación: Facultad de Ingeniería, Universidad San Ignacio de Loyola

1. Resumen

El presente trabajo aborda la generación de imágenes sintéticas de prendas de vestir utilizando el dataset Fashion-MNIST y una arquitectura de Deep Convolutional GAN (DCGAN) en tres configuraciones distintas. El objetivo fue evaluar cómo el tamaño del espacio latente (z_dim) y el tamaño de batch influyen en la calidad de las imágenes generadas.

Para la comparación se utilizaron las métricas Frechet Inception Distance (FID), Kernel Inception Distance (KID) e Inception Score (IS), además de un análisis cualitativo mediante rejillas de imágenes (inicio, mitad y final del entrenamiento) y una comparación entre una imagen generada y la real más parecida.

Los resultados muestran que el modelo B (z_dim=50) alcanzó el mejor FID, mientras que el modelo A (z_dim=100) logró un mejor IS, destacando el rol del tamaño del espacio latente en el balance entre fidelidad y diversidad. Estos hallazgos permiten reflexionar sobre el diseño de GANs ligeras y sus posibles aplicaciones en moda digital y generación de datos sintéticos.

2. Introducción

La generación de imágenes mediante redes adversarias generativas (GANs) se ha convertido en un área fundamental del deep learning, con aplicaciones que incluyen desde creación de contenido hasta el aumento de datos para mejorar modelos supervisados [1]. En particular, en el ámbito de la moda digital, la capacidad de generar prendas sintéticas realistas puede contribuir a acelerar procesos de diseño, pruebas virtuales y sistemas de recomendación [2].

Este proyecto tiene como objetivo comparar tres configuraciones de una DCGAN entrenada sobre el dataset Fashion-MNIST, que contiene imágenes en escala de grises de diez categorías de ropa [3]. Dado que la evaluación de GANs difiere de las tareas clásicas de clasificación, se emplean métricas específicas para medir la calidad y diversidad de las imágenes generadas, como FID, KID e IS, acompañadas de evaluaciones visuales a lo largo del entrenamiento [4] [5]

La investigación busca responder a la siguiente pregunta central:

¿Cómo influyen las variaciones en el espacio latente y el batch size en la calidad de las imágenes generadas por una DCGAN ligera?

3. Datos

El conjunto de datos utilizado en este trabajo proviene de Fashion-MNIST, desarrollado por Zalando Research (disponible en https://github.com/zalandoresearch/fashion-mnist). Este dataset contiene un total de 70,000 imágenes en escala de grises con una resolución de 28×28 píxeles, distribuidas en 10 clases

distintas de prendas de vestir, como camisetas, zapatos y bolsos. Las imágenes están divididas en 60,000 muestras para entrenamiento y 10,000 para prueba.

Ello se realizó mediante el siguiente código:

```
# Ruta donde se almacenarán los datos
data_root = "data"

# Carga train/test sin transformaciones
train_data = datasets.FashionMNIST(root=data_root, train=True,
download=True)
test_data  = datasets.FashionMNIST(root=data_root, train=False,
download=True)

print(f"Imágenes de entrenamiento: {len(train_data)}")
print(f"Imágenes de prueba: {len(test_data)}")
```

```
100%|██████████| 26.4M/26.4M [00:00<00:00, 112MB/s]
100%|██████████| 29.5k/29.5k [00:00<00:00, 4.15MB/s]
100%|██████████| 4.42M/4.42M [00:00<00:00, 54.9MB/s]
100%|██████████| 5.15k/5.15k [00:00<00:00, 30.7MB/s]Imágenes de entrenamiento: 60000
Imágenes de prueba: 10000
```

4. Metodología

a. Ruta Elegida y Modelos

Para este trabajo se eligió la ruta de generación de imágenes mediante Redes Adversarias Generativas (GANs). En especial, se implementó una Deep Convolutional GAN (DCGAN), una arquitectura ampliamente utilizada por su efectividad y simplicidad en visión por computadora, compuesta por:

- Generador (G): transforma un vector de ruido aleatorio z en una imagen sintética.
- Discriminador (D): distingue entre imágenes reales y generadas [6].

Se evaluaron tres configuraciones del modelo (A, B y C), diferenciadas por el tamaño del espacio latente y el batch size:

- Modelo A: z_dim=100, batch size = 128.
- Modelo B: z_dim=50, batch size = 64.
- Modelo C: z_dim=200, batch size = 128.

El espacio latente es un hiperparámetro crítico: estudios recientes muestran que dimensiones demasiado pequeñas pueden limitar la diversidad generada, mientras que dimensiones excesivamente grandes pueden afectar la estabilidad del entrenamiento [7] [8].

El resto de hiperparámetros (épocas, tasa de aprendizaje, optimizadores) se mantuvo constante para asegurar una comparación justa.

b. Preprocesamiento y configuración de entrenamiento

El dataset Fashion-MNIST fue cargado desde torchvision.datasets, conteniendo imágenes en escala de grises de 28×28 píxeles correspondientes a 10 categorías de ropa.

- **Normalización:** las imágenes fueron escaladas al rango [-1, 1] para ajustarse a la salida de la función *tanh* del generador.
- **Particiones:** se usó el conjunto de entrenamiento (60,000 imágenes) para entrenar los modelos, mientras que el conjunto de prueba (10,000 imágenes) fue reservado solo para la comparación cualitativa en la etapa de inferencia.
- **Batching:** se aplicó el tamaño de batch definido en cada configuración (64 o 128).
- **Optimizadores:** se utilizó Adam con tasa de aprendizaje lr=0.0002 y parámetros $\beta1=0.5$, $\beta2=0.999$.
- **Épocas:** cada modelo fue entrenado durante 30 épocas.
- **Registro de entrenamiento:** se guardaron las pérdidas de generador y discriminador en cada época, así como imágenes de muestra en tres momentos: inicio, mitad y final del entrenamiento.

c. Métricas seleccionadas

Para evaluar el desempeño de los modelos, se usaron métricas específicas para GANs, aplicadas en el esquema eval-real (comparación entre imágenes reales y generadas)**:**

1. **Frechet Inception Distance (FID):** calcula la distancia entre las distribuciones de características extraídas por una red Inception-v3 desde imágenes reales vs. generadas, usando la distancia de Fréchet sobre medias y covarianzas. Menor es mejor [9].
2. **Kernel Inception Distance (KID):** alternativa a FID basada en la distancia de máxima discrepancia media (MMD), que no requiere estimación gaussiana y es insesgada, lo que la hace más confiable en datasets pequeños [10] [11].
3. **Inception Score (IS):** mide la calidad y diversidad de las imágenes generadas. Un puntaje más alto indica que las imágenes son diversas y reconocibles. En este trabajo, se reportaron dos valores:

    - IS_mean: promedio del puntaje en varios lotes de imágenes.
    - IS_std: desviación estándar, que refleja la estabilidad del generador frente a variabilidad en los resultados [12]

Las métricas se calcularon en el conjunto de evaluación (eval–real): comparando lotes de imágenes reales del dataset con lotes generados por cada modelo para capturar objetivamente calidad (FID, KID) y diversidad/claridad (IS).

5. Resultados

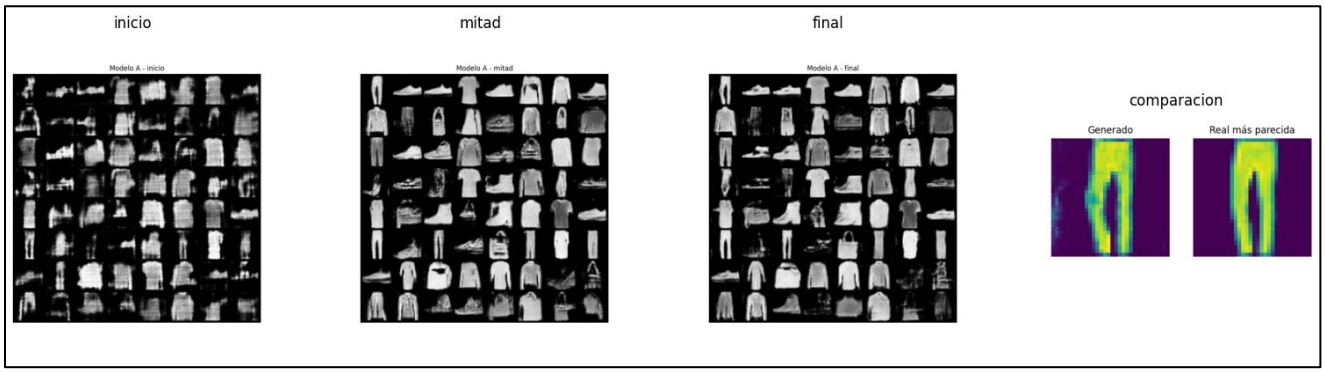
a. Evidencias de inferencia

Durante el entrenamiento se generaron rejillas 8×8 de imágenes con el mismo ruido en tres momentos clave: inicio, mitad y final de entrenamiento. Esto permitió visualizar la evolución de la calidad y nitidez de las prendas sintéticas. Se observó que:

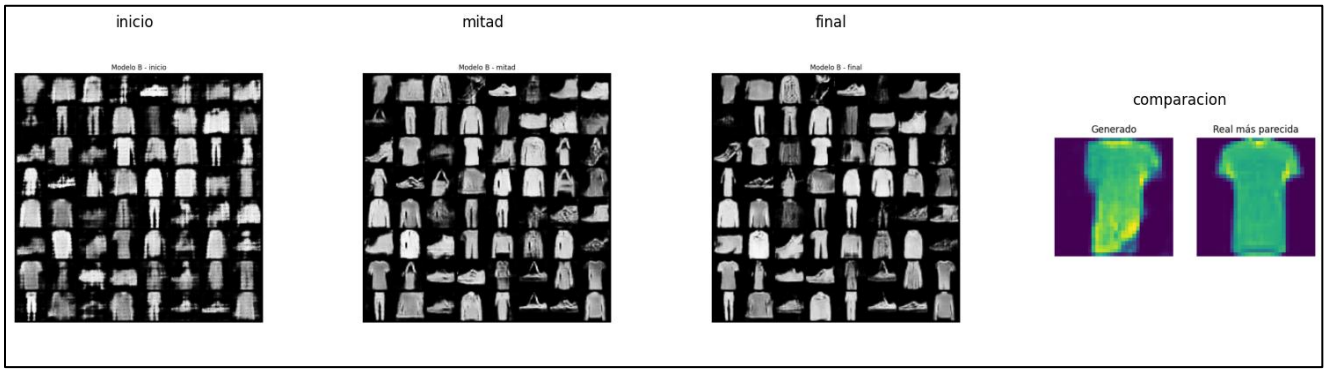
- En las primeras épocas, las imágenes eran mayormente ruido sin forma reconocible.
- Hacia la mitad del entrenamiento, comenzaron a aparecer siluetas básicas de prendas como camisetas y zapatos.
- Al final, las muestras mostraban mayor definición, aunque con limitaciones en detalles finos.

Adicionalmente, se realizó la comparación "imagen generada vs. real más parecida", donde se evidenció que las imágenes sintéticas lograban aproximarse en forma general a las reales, aunque sin replicar detalles de textura o bordes nítidos.

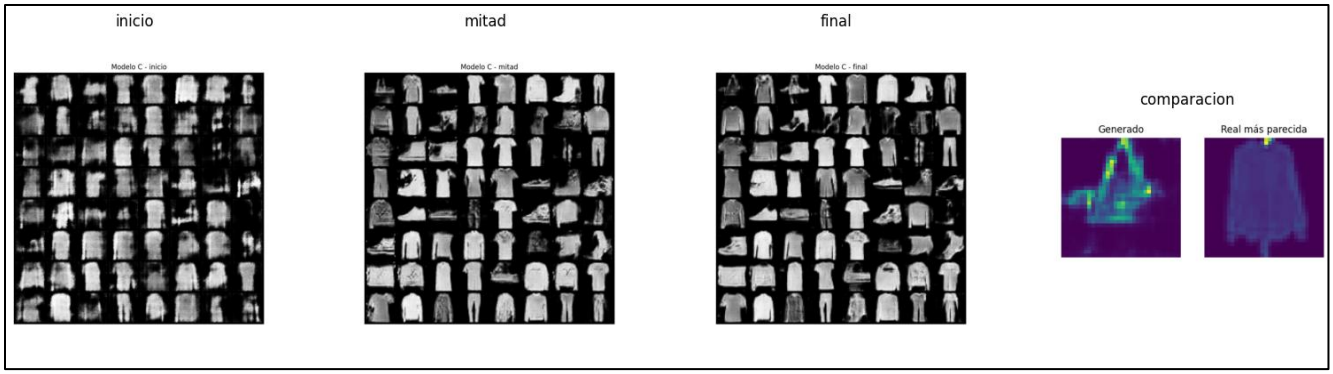
i. Configuración A



ii. Configuración B



iii. Configuración C



b. Gráficos por métrica

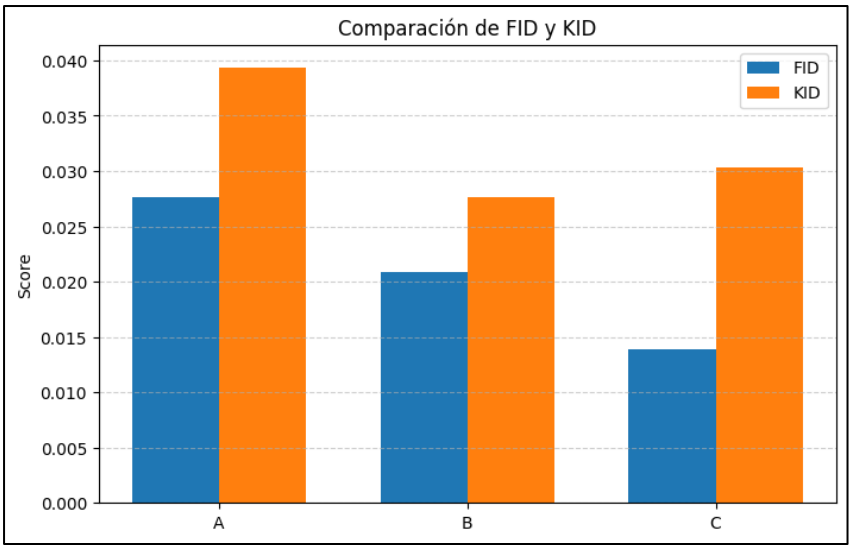
Se evaluaron las tres configuraciones (A, B y C) con las métricas FID, KID e Inception Score (mean ± std).

i. Comparación FID, KID

En la comparación de FID **y** KID**,** se observa que:

- La configuración C (z_dim=200) obtuvo el mejor **FID (0.0139)**, lo que indica que sus imágenes sintéticas fueron estadísticamente más cercanas a las reales.
- La configuración B (z_dim=50, batch=64) destacó en **KID (0.0276)**, siendo la que generó imágenes más consistentes en datasets pequeños.
- La configuración A (z_dim=100) alcanzó valores intermedios en ambas métricas.

Esto sugiere que la elección de parámetros impacta de forma diferenciada según la métrica: un espacio latente grande (C) puede mejorar la similitud global (FID), mientras que un espacio más pequeño (B) ayuda a la consistencia (KID).

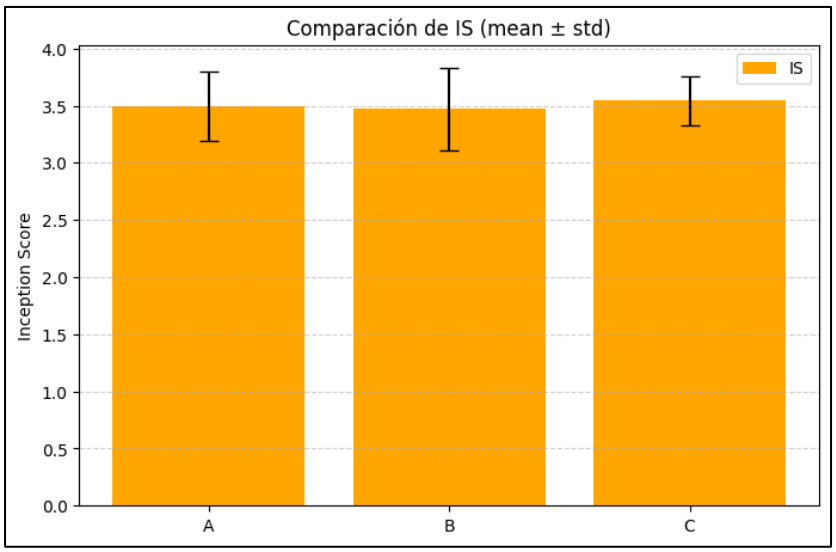


ii. Comparación IS

En términos de Inception Score (IS):

- La configuración C nuevamente fue la mejor, con un IS_mean de 3.54 ± 0.22, lo que refleja mayor diversidad y claridad en las imágenes generadas.
- Las configuraciones A (3.50 ± 0.30) y B (3.47 ± 0.36) obtuvieron resultados similares, aunque con mayor variabilidad (desviación estándar más alta).

Esto refuerza que la configuración C genera imágenes más variadas y consistentes.



iii. Curvas de pérdida del generador y discriminador

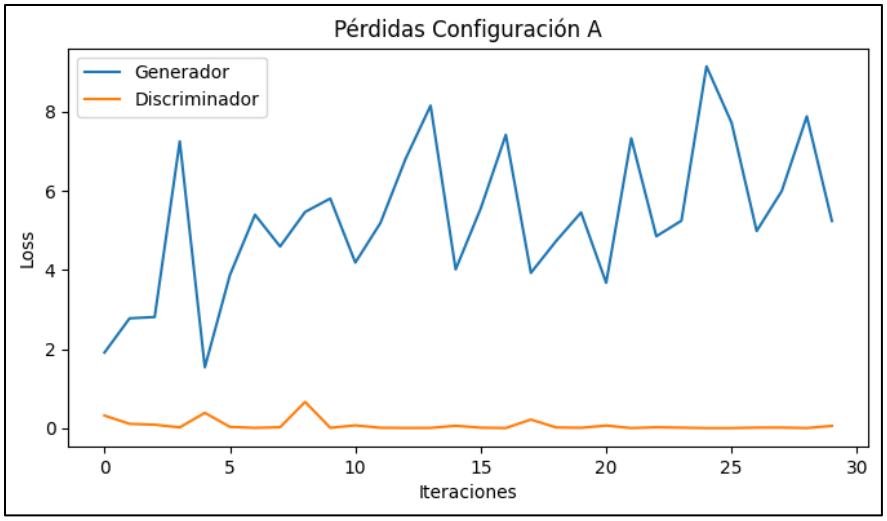
Al analizar las curvas de pérdida en cada configuración:

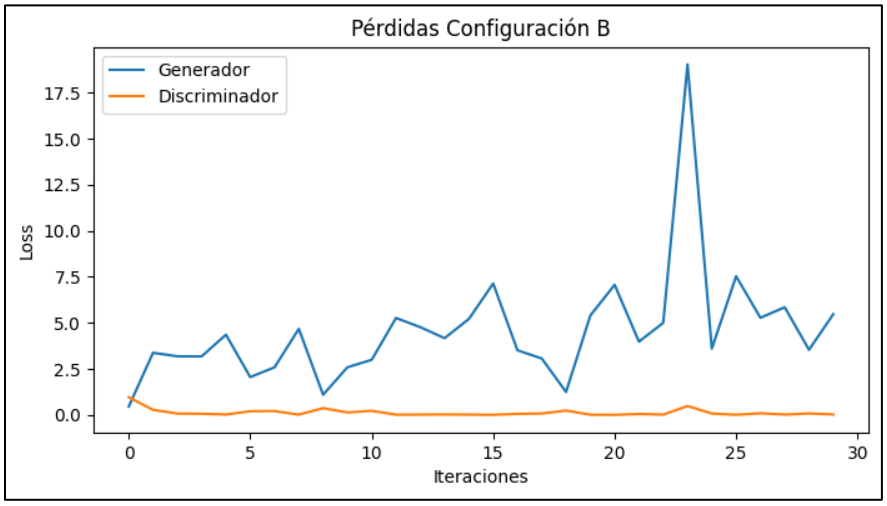
En todas, el discriminador mantiene valores bajos y relativamente estables, lo que indica que rápidamente aprendió a distinguir entre imágenes reales y falsas.

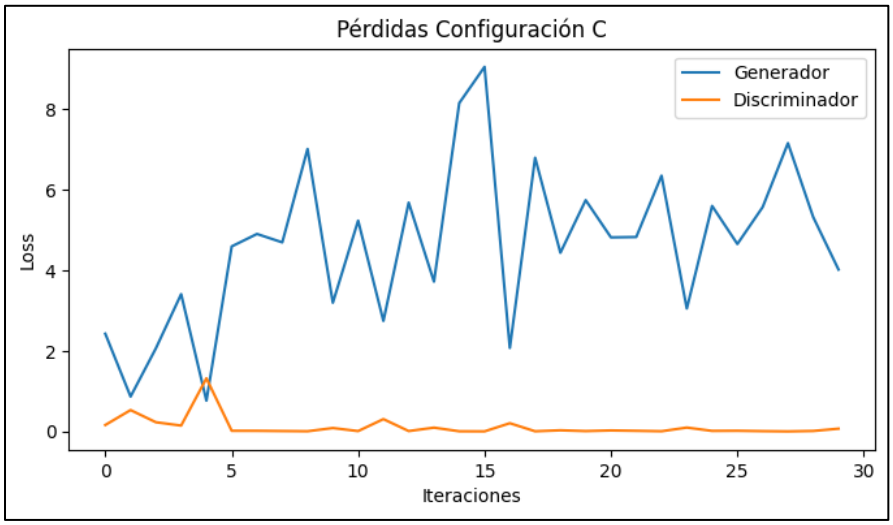
El generador presenta más inestabilidad con picos y caídas.

- Configuración A: el mayor pico ocurrió cerca de la época 25, mostrando una oscilación marcada.
- Configuración B: el generador se mantuvo bajo 7.5 hasta la época 23, donde alcanzó un pico mayor a 17.5, lo que refleja alta inestabilidad puntual.
- Configuración C: mostró subidas y bajadas que no superaron 6.5, con un pico en torno a la época 15 (~8.0). A pesar de ello, sus valores se mantuvieron más agrupados en comparación con A y B.

En resumen, la configuración C fue la más estable en las pérdidas, lo que está en consonancia con sus resultados en métricas.





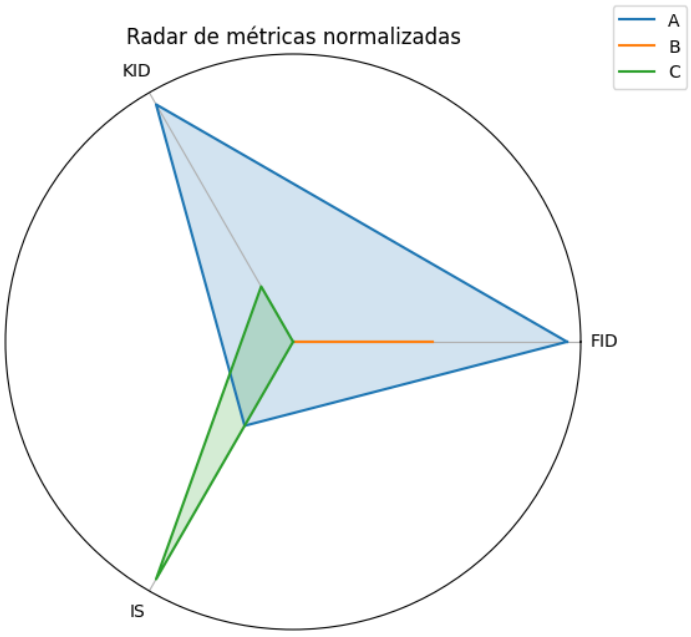


iv. Radar de métricas normalizadas

El gráfico radar de métricas normalizadas mostró un patrón claro:

- La configuración C domina tanto en FID como en IS, además de mantener un desempeño sólido en KID.
- La configuración B es competitiva en KID, pero rezagada en FID e IS.
- La configuración A queda en una posición intermedia en todas las métricas, sin destacar en ninguna.

Este análisis consolidado permite concluir que, si bien cada configuración tiene fortalezas, la configuración C es la más equilibrada y la mejor opción global para la generación de imágenes en Fashion-MNIST.



c. Tabla Comparativa

Configuración	FID	KID	IS_mean	IS_std	Pérdida G (final)	Pérdida D (final)	Observaciones
A (z=100)	0.027636	0.039390	3.496917	0.303776	5.2393	0.0568	Generador con picos pronunciados (máx. ≈9.15 en época 25). Discriminador estable y bajo. Buen balance entre FID/KID.

B (z=50)	0.020862	0.027615	3.471461	0.363097	5.4666	0.0190	Generador inestable con pico extremo (≈19 en época 24). Aun así logra buen FID/KID. Discriminador consistente.
C (z=200)	0.013888	0.030357	3.543423	0.216749	4.0203	0.0688	Generador más estable, sin picos extremos (máx. ≈9). Mejor IS y menor FID. Discriminador con pérdidas algo más altas que A y B.

Comentario: La tabla refleja un desempeño diferenciado entre las configuraciones. La configuración C (z=200) alcanzó el mejor resultado en FID (0.0139) e Inception Score (3.54 ± 0.22), mostrando imágenes más realistas y diversas con mayor estabilidad (menor desviación estándar). Además, presentó un comportamiento más estable en las curvas de pérdida, sin picos extremos en el generador (máx. ≈9), lo que refuerza su consistencia.

La configuración B (z=50) destacó en KID (0.0276), lo que sugiere mayor consistencia en datasets pequeños, aunque no superó a C en las demás métricas. Sin embargo, su entrenamiento fue más inestable: el generador presentó un pico muy elevado de pérdida (≈19 en la época 24), lo que refleja mayor volatilidad.

La configuración A (z=100) se mantuvo en valores intermedios en todas las métricas. Su generador mostró picos pronunciados (máx. ≈9.15 en la época 25), mientras que el discriminador se mantuvo con pérdidas bajas y estables, indicando un aprendizaje más controlado en esa parte de la arquitectura.

En conjunto, el análisis sugiere que la configuración C es la más balanceada y la opción óptima, mientras que B puede ser preferida en escenarios donde la robustez en KID sea prioritaria, a costa de mayor inestabilidad en el generador.

Código disponible en: https://github.com/eviediaz/GAN_Experiment

6. Conclusiones

El presente trabajo exploró la generación de imágenes sintéticas de ropa utilizando el dataset Fashion-MNIST con tres configuraciones de DCGAN que variaron principalmente en la dimensión del espacio latente (*z_dim*) y el tamaño de batch. El análisis, basado en métricas cuantitativas **(**FID, KID e IS**)** y evaluaciones cualitativas (rejillas de imágenes y comparaciones "real vs. generado"), permitió extraer varias conclusiones:

1. **Influencia del espacio latente:** se evidenció que un espacio latente más amplio **(**configuración C, z=200**)** favorece tanto la fidelidad como la diversidad de las imágenes, logrando los mejores valores en FID (0.0139) e IS (3.54 ± 0.22)**.**
2. **Robustez en conjuntos pequeños:** la configuración B (z=50, batch=64) obtuvo el mejor KID (0.0276)**,** lo que la hace más adecuada en escenarios donde los datos reales disponibles son limitados.
3. **Estabilidad de entrenamiento:** aunque todas las configuraciones presentaron oscilaciones típicas en la pérdida del generador, la configuración **C** fue la más estable, en consonancia con sus resultados en métricas.
4. **Valor de las evaluaciones cualitativas:** las rejillas de imágenes confirmaron que el entrenamiento progresivo mejora significativamente la calidad visual, pasando de ruido inicial a prendas reconocibles.

En conjunto, se concluye que la configuración C es la opción óptima para la tarea de generación de imágenes en Fashion-MNIST, al equilibrar realismo, diversidad y estabilidad. No obstante, el hecho de que la configuración B obtuviera un KID superior muestra que el ajuste de hiperparámetros debe adaptarse al contexto de aplicación.

Lecciones aprendidas: la comparación resaltó la importancia de evaluar GANs con múltiples métricas y no depender de un solo indicador, ya que cada métrica captura aspectos distintos de calidad y diversidad. Asimismo, se evidenció el rol crítico del preprocesamiento y la necesidad de visualizar las salidas durante el entrenamiento para una interpretación más completa del desempeño.

7. Limitaciones y Ética

El desarrollo y uso de modelos generativos como las GANs presenta tanto limitaciones técnicas como consideraciones éticas relevantes que deben ser discutidas:

a. <u>Posibles sesgos en los datos</u>

El uso del dataset Fashion-MNIST, compuesto únicamente por imágenes a escala de grises de baja resolución (28 × 28), limita la representatividad de la moda real y puede introducir sesgos en el entrenamiento del modelo [3].

Además, se han identificado numerosos errores de etiquetado en el conjunto, lo que afecta la calidad del entrenamiento y reproduce errores inherentes al dataset [13].

b. Riesgos de uso

Las GANs tienen un gran potencial para genera contenido creíble, pero también pueden facilitar la creación de deepfakes o imágenes manipuladas con fines engañosos o malintencionados [14].

Existe asimismo el riesgo de propagación de sesgos presentes en los datos de entrenamiento, lo que puede derivar en salidas estereotipadas o discriminatorias [14].

El uso de datos sintéticos en sectores regulados, como finanzas, puede generar situaciones de manipulación o falta de supervisión ética en la generación de información falsa [15].

c. Mejoras futuras

Para mitigar las limitaciones técnicas y éticas, se proponen varias líneas de mejora:

- Datasets más diversos y realistas: incorporar bases de datos de ropa en color, con mayor resolución y variedad cultural.
- Arquitecturas más robustas: explorar variantes como WGAN-GP, Conditional GANs o modelos de difusión, que suelen ser más estables y producen mayor calidad visual [16].
- Mecanismos de control ético: incluir trazabilidad de imágenes generadas y mecanismos de marca de agua que permitan identificar contenido sintético.
- Aplicar técnicas de mitigación de sesgo, como repfair-GAN, que busca asegurar justicia representacional en modelos generativos mediante recorte de gradientes por grupo [17].
- Evaluaciones integrales: complementar las métricas automáticas (FID, KID, IS) con evaluaciones humanas que consideren realismo, creatividad y representatividad.

8. Referencias

[1] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, y A. A. Bharath, «Generative Adversarial Networks: An Overview», *IEEE Signal Process. Mag.*, vol. 35, n.º 1, pp. 53-65, ene. 2018, doi: 10.1109/MSP.2017.2765202.

[2] A. Mumuni, F. Mumuni, y N. K. Gerrar, «A Survey of Synthetic Data Augmentation Methods in Machine Vision», *Mach. Intell. Res.*, vol. 21, n.º 5, pp. 831-869, oct. 2024, doi: 10.1007/s11633-022-1411-7.

[3] H. Xiao, K. Rasul, y R. Vollgraf, «Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms», 15 de septiembre de 2017, *arXiv*: arXiv:1708.07747. doi: 10.48550/arXiv.1708.07747.

[4] A. Borji, «Pros and cons of GAN evaluation measures: New developments», *Computer Vision and Image Understanding*, vol. 215, p. 103329, ene. 2022, doi: 10.1016/j.cviu.2021.103329.

[5] S. Carrasco, «On the evaluation of Generative Adversarial Networks», ene. 2025. [En línea]. Disponible en: https://towardsdatascience.com/on-the-evaluation-of-generative-adversarial-networks-b056ddcdfd3a/

[6] A.-M. Simion, Şerban Radu, y A. M. Florea, «A Review of Generative Adversarial Networks for Computer Vision Tasks», *Electronics*, vol. 13, n.º 4, p. 713, feb. 2024, doi: 10.3390/electronics13040713.

[7] İ. Ataş, «The Effect of Latent Space Vector on Generating Animal Faces in Deep Convolutional GAN: An Analysis», *DÜMF MD*, mar. 2024, doi: 10.24012/dumf.1393797.

[8] I. Marin, S. Gotovac, M. Russo, y D. Božić-Štulić, «The Effect of Latent Space Dimension on the Quality of Synthesized Human Face Images», *JCOMSS*, vol. 17, n.º 2, pp. 124-133, 2021, doi: 10.24138/jcomss-2021-0035.

[9] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, y S. Hochreiter, «GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium», 2017, doi: 10.48550/ARXIV.1706.08500.

[10] M. Bińkowski, D. J. Sutherland, M. Arbel, y A. Gretton, «Demystifying MMD GANs», 2018, *arXiv*. doi: 10.48550/ARXIV.1801.01401.

[11] Arthur Gretton, «GANs with Integral Probability Metrics: Some results and conjectures», Gatsby Computational Neuroscience Unit, 2019. [En línea]. Disponible en: https://www.gatsby.ucl.ac.uk/~gretton/papers/montreal19.pdf

[12] A. Brock, J. Donahue, y K. Simonyan, «Large Scale GAN Training for High Fidelity Natural Image Synthesis», 2018, *arXiv*. doi: 10.48550/ARXIV.1809.11096.

[13] G. Tata y C. Mauck, «The Fashion MNIST Dataset (cited in 2,200+ papers) contains Hundreds of Miscategorized Items», Cleanlab, 2023. [En línea]. Disponible en: https://cleanlab.ai/blog/csa/csa-4/

[14] Wikipedia contributors, «Generative adversarial network», *Wikipedia*. [En línea]. Disponible en: https://en.wikipedia.org/wiki/Generative_adversarial_network

[15] A. Naidu, «Ethical Implications of Using GANs in the Financial Sector: Balancing Innovation with Security», *IJMRGE*, vol. 2, n.º 5, pp. 474-477, 2021, doi: 10.54660/.IJMRGE.2021.2.5.474-477.

[16] U. Sirisha, C. K. Kumar, S. C. Narahari, y P. N. Srinivasu, «An Iterative PRISMA Review of GAN Models for Image Processing, Medical Diagnosis, and Network Security», *CMC*, vol. 82, n.º 2, pp. 1757-1810, 2025, doi: 10.32604/cmc.2024.059715.

[17] P. J. Kenfack, K. Sabbagh, A. R. Rivera, y A. Khan, «RepFair-GAN: Mitigating Representation Bias in GANs Using Gradient Clipping», 2022, *arXiv*. doi: 10.48550/ARXIV.2207.10653.