



Instituto Politécnico Nacional
Unidad Profesional Interdisciplinaria de Ingeniería
Campus Tlaxcala



Reporte Técnico
Cálculo de Similitud entre Imágenes
Artísticas y Generadas por Stable Diffusion
utilizando Redes Convolucionales
Siamesas
No. TT2025-1_IA-008

Presenta(n):

Diego Castro Elvira, Ingeniería en Inteligencia Artificial
Navil Pineda Rugerio, Ingeniería en Inteligencia Artificial

Asesores:

Dr. Ricardo Ramos Aguilar

Dr. Jesús García Ramírez

Tlaxcala, Tlaxcala, a **05 de diciembre de 2024**

"La Técnica al Servicio de la Patria"

AGRADECIMIENTOS

Diego Castro Elvira:

Agradezco a la vida, por darme la fuerza para continuar, la sabiduría para aprender y la resiliencia para superar los desafíos. Este trabajo es el reflejo de cada pequeño paso, de cada amanecer que me recordó que siempre hay una nueva oportunidad para crecer y avanzar.

Navil Pineda Rugerio:

Agradezco a Dios, por ser mi guía y mi fuente de sabiduría en todo momento. A mis padres, Alma Delia Rugerio Díaz y Abraham Pineda García, por ser mi inspiración, creer en mí y apoyarme en cada paso de mi vida. A mis hermanos, Alma Karen y Abraham, por su confianza y palabras de ánimo en los momentos más difíciles. A mis asesores, Dr. Jesús García Ramírez y Dr. Ricardo Ramos Aguilar, por todo el tiempo dedicado en el desarrollo de este proyecto, orientación y valiosas enseñanzas. A mi amigo y compañero de Trabajo Terminal, Diego Castro Elvira, con quien ha sido un honor trabajar, ya que con su esfuerzo y apoyo constante llevamos a cabo este proyecto.

RESUMEN

La Inteligencia Artificial generativa ha permitido crear obras más complejas estéticamente, comparables a las creadas por artistas humanos, lo que implica nuevas dificultades para evaluar la originalidad y autenticidad de estas. Si bien existen investigaciones y herramientas digitales dedicadas al análisis comparativo de obras artísticas, hasta el momento, no se han desarrollado modelos que permitan evaluar la similitud entre obras originales y generadas mediante técnicas de image-to-image que intentan emular el estilo de creaciones humanas.

El presente proyecto tiene como objetivo calcular la similitud entre imágenes artísticas originales y aquellas generadas por el modelo Stable Diffusion XL Refiner 1.0, utilizando redes neuronales convolucionales siamesas con *fine-tuning* a un modelo preentrenado ResNet50.

Para llevar a cabo este proceso, se crea un conjunto de datos compuesto por imágenes provenientes de WikiArt, junto con sus correspondientes versiones generadas por IA. A continuación, se seleccionan y extraen una serie de descriptores visuales, los cuales se utilizan para calcular valores de similitud entre las imágenes originales y las generadas por IA. Estos valores obtenidos a partir de los descriptores son luego comparados con los resultados producidos por la red siamesa entrenada. El rendimiento del modelo se evalúa utilizando métricas de regresión estándar, como el Error Cuadrático Medio (MSE) y el Error Absoluto Medio (MAE), con el propósito de obtener un valor cuantitativo que refleje la similitud entre las imágenes analizadas.

Términos/Palabras Clave

- Descriptores de imagen
- Redes convolucionales siamesas
- Similitud de imágenes
- *Stable Diffusion*

ÍNDICE GENERAL

1. Introducción	1
1.1. Antecedentes	1
1.2. Planteamiento del Problema	1
1.2.1. Definición del Problema	1
1.2.2. Objetivos	2
1.2.3. Justificación	2
1.3. Hipótesis	3
1.4. Aportación Científica y/o Tecnológica	3
1.5. Organización del Documento	3
2. Marco Teórico	5
2.1. Redes Neuronales Convolucionales	5
2.1.1. Arquitectura de las CNNs	5
2.1.2. <i>Fine-tuning</i>	7
2.1.3. ResNet	7
2.2. Redes Siamesas	8
2.2.1. Arquitectura de las redes siamesas	8
2.3. Generación de Imágenes mediante Modelos Generativos	8
2.3.1. Modelos generativos de imágenes y su evolución	8
2.3.2. Modelos Generativos de Difusión	9
2.3.3. <i>Stable Diffusion</i>	11
2.4. Descriptores de imágenes	15
2.4.1. Tipos de descriptores	15
2.5. Estado del Arte	17
3. Metodología	21
3.1. Investigación preliminar.	21
3.2. Creación del conjunto de datos.	22
3.3. Selección de descriptores de imágenes.	23
3.4. Generación de valores de similitud.	23
3.5. <i>Fine-tuning</i> del modelo ResNet50.	24
3.6. Implementación y entrenamiento de la Red Siamesa.	24
3.7. Evaluación.	24

CAPÍTULO 1

INTRODUCCIÓN

El uso de la inteligencia artificial (IA) en la creación artística ha crecido considerablemente, planteando interrogantes sobre la originalidad de las obras y su posible similitud con creaciones humanas, mayormente en el arte digital, donde modelos como *Stable Diffusion* [51] permiten la generación de imágenes realistas y estilizadas. Sin embargo, la falta de herramientas que permitan evaluar la autenticidad de estas obras ha generado la necesidad de desarrollar métodos que ayuden a identificar posibles similitudes en el arte generado por IA.

Este proyecto surge del interés tanto académico como profesional por mejorar los procesos de autenticación de obras artísticas digitales, facilitando la detección de similitudes entre imágenes originales y aquellas generadas por IA. Para ello, se ha optado por emplear un enfoque basado en redes neuronales convolucionales siamesas, capaces de comparar imágenes a partir de descriptores visuales y medir su grado de similitud.

La metodología sigue un proceso incremental que incluye la creación de un conjunto de datos, *fine-tuning* de un modelo pre-entrenado y la implementación de una red siamesa para el análisis de las imágenes. A través de este trabajo, se busca aportar una herramienta para la valoración de la originalidad en el arte generado por IA, apoyando a artistas y la comunidad de investigación en la identificación de posibles similitudes.

1.1. Antecedentes

1.2. Planteamiento del Problema

1.2.1. Definición del Problema

La IA generativa ha permitido crear obras más complejas estéticamente, comparables a las creadas por artistas humanos, lo que implica nuevas dificultades para evaluar la originalidad y autenticidad de estas. Un aspecto clave al comparar imágenes es la identificación de detalles que, debido a la

1.2. PLANTEAMIENTO DEL PROBLEMA

complejidad de las pinturas artísticas, varían en características perceptibles al ojo humano, como la composición, el color o la iluminación. Sin embargo, comparar estos elementos de manera conjunta se vuelve un desafío considerable.

Aunque ya existen estudios y herramientas de software centrados en la comparación y el análisis del arte digital, persiste una brecha en la investigación, no se han desarrollado modelos capaces de evaluar la similitud entre las obras originales y las imágenes generadas por técnicas de *image-to-image*, que buscan replicar el estilo de obras creadas por humanos.

Además, la mayoría de los estudios se centran en estilos específicos, sin abarcar una amplia gama de géneros, y no existen conjuntos de datos que incluyan tanto obras originales como generadas por IA, lo que limita el desarrollo de modelos más robustos para el análisis comparativo.

1.2.2. Objetivos

Objetivo General

Construir un modelo para calcular la similitud entre imágenes de pinturas artísticas y aquellas generadas por *Stable Diffusion XL Refiner 1.0* mediante el uso de descriptores de imagen y una red convolucional siamesa entrenada con fine-tuning a una ResNet50.

Objetivos Específicos

- Crear un conjunto de datos con imágenes originales de WikiArt [50] y sus versiones generadas por *Stable Diffusion XL Refiner 1.0*.
- Seleccionar y extraer los descriptores de las imágenes del dataset de WikiArt y de las imágenes generadas por *Stable Diffusion* utilizando técnicas de procesamiento de imágenes.
- Diseñar y entrenar una red convolucional siamesa para calcular la similitud entre las imágenes del dataset WikiArt y las imágenes generadas por *Stable Diffusion*.

1.2.3. Justificación

Debido a la falta de modelos especializados que aborden la generación y evaluación de similitud en pinturas de diversos géneros artísticos, este proyecto propone un nuevo enfoque al combinar la generación de imágenes similares mediante técnicas de *image-to-image* con la evaluación de su similitud, utilizando redes convolucionales siamesas, y plantea desarrollar una compilación de un conjunto de datos que incluya obras originales y sus versiones generadas por IA en diversos estilos, ya que a diferencia de estudios previos, este proyecto propone una gama de géneros artísticos más diversa.

Con ello, se busca abordar el desafío de la autenticación y originalidad en el arte digital, beneficiando a artistas, galerías, plataformas en línea y la comunidad de investigación. Su viabilidad está asegurada por el uso de recursos computacionales accesibles y un cronograma adecuado para su desarrollo y prueba.

1.3. Hipótesis

Las redes neuronales siamesas permiten medir cuantitativamente el nivel de similitud compositiva y estilística entre pinturas artísticas originales y sus contrapartes generadas por *Stable Diffusion XL Refiner 1.0*, mediante la identificación y comparación de sus patrones visuales o semánticos.

1.4. Aportación Científica y/o Tecnológica

Los aportes esperados son los siguientes:

1. Desarrollo de un conjunto de datos de imágenes artísticas y generadas por *Stable Diffusion*: El proyecto generará y pondrá a disposición un conjunto de datos con imágenes de WikiArt y sus correspondientes versiones creadas con *Stable Diffusion XL Refiner 1.0*, lo cual servirá como recurso para investigaciones futuras en el análisis de similitud entre imágenes originales y generadas.
2. Implementación de un modelo de red siamesa especializada: Al finalizar el proyecto, se contará con un modelo de red siamesa, ajustado mediante *fine-tuning* de una ResNet50, diseñado específicamente para la comparación y cálculo de similitud entre obras de arte y sus versiones generadas. Este modelo puede aplicarse en investigaciones futuras o en sistemas de recomendación de arte.
3. Contribución a la comprensión de descriptores de imagen para análisis de arte: Al seleccionar y analizar diferentes descriptores de imagen, el proyecto aportará al conocimiento sobre cuáles características visuales son más relevantes al comparar obras artísticas y versiones generadas, ayudando en el desarrollo de futuros modelos de análisis de arte.

1.5. Organización del Documento

En la introducción se establece el contexto general y la importancia de la investigación. Comienza con los antecedentes para situar el proyecto dentro de un marco teórico relevante, seguido del planteamiento del problema, que define el enfoque y objetivos de estudio. La justificación resalta

1.5. ORGANIZACIÓN DEL DOCUMENTO

el valor que esta investigación aporta en el ámbito de la inteligencia artificial y el análisis de imágenes, mientras que los objetivos detallan las metas específicas que se desean lograr. Además, la hipótesis sugiere el resultado esperado; y la aportación científica y tecnológica proyecta los beneficios prácticos que se esperan obtener, como publicaciones, desarrollos de software y nuevos conocimientos en el área de comparación de imágenes.

El marco teórico profundiza en los conceptos técnicos y metodológicos clave que sustentan el proyecto. Inicia con una exploración de las redes neuronales convolucionales, esenciales para la implementación del modelo siamesa, y discute el *fine-tuning* en modelos preentrenados, centrándose en *ResNet50*, que será ajustado para el análisis de similitud entre imágenes. También se abordan las redes siamesas desde su origen y arquitectura, dado su rol en la comparación de similitud. Luego se exploran los modelos generativos, en especial *Stable Diffusion*, para entender los procesos de creación de imágenes sintéticas que serán comparadas con obras de arte originales. Finalmente, los descriptores de imágenes y el estado del arte aportan una revisión de investigaciones previas, ofreciendo un panorama de los avances y las limitaciones actuales en este campo.

La metodología describe el enfoque experimental de la investigación, cubriendo desde la creación del conjunto de datos hasta el diseño y ajuste de la red siamesa, así como el *fine-tunning* de *ResNet50*. En esta sección se detallan también los métodos de evaluación del modelo, de manera que el proceso pueda replicarse y ofrezca claridad sobre cada paso seguido en la investigación.

CAPÍTULO 2

MARCO TEÓRICO

2.1. Redes Neuronales Convolucionales

Las redes neuronales convolucionales (*Convolutional Neural Network*, CNN) son una clase de algoritmos de aprendizaje profundo diseñados para procesar datos estructurados en forma de cuadrícula, sobre todo imágenes. Aprovechan una arquitectura en capas para aprender de forma automática y adaptativa jerarquías espaciales de características, lo que las hace especialmente eficaces para tareas de visión por computadora, como la clasificación de imágenes, la detección de objetos y la segmentación semántica.

El origen de las CNNs se remonta a los años 50 con la introducción del perceptrón por Frank Rosenblatt, que sentó las bases para el reconocimiento de patrones [20]. En los años 80, Kunihiko Fukushima desarrolló el "neocognitron", una forma temprana de CNN inspirada en arquitecturas neuronales biológicas. Este modelo fue refinado en los años 90 con la creación de LeNet-5, una CNN diseñada para el reconocimiento de dígitos manuscritos, lo que demostró su utilidad práctica y fomentó más investigación en el campo [23].

2.1.1. Arquitectura de las CNNs

La arquitectura de las CNNs está diseñada para aprender automáticamente jerarquías espaciales de características a partir de imágenes de entrada, su estructura incluye:

1. Capas Convolucionales: Esta capa está compuesta por un conjunto de núcleos de convolución. Su función es deslizar una ventana de tamaño fijo sobre el mapa de características, extrayendo mosaicos de características en diferentes posiciones. Posteriormente, se realiza el producto tensorial entre cada mosaico y el núcleo de convolución, reorganizando el espacio vectorial para generar un nuevo tensor.
2. Funciones de Activación: Las funciones de activación permiten que las

2.1. REDES NEURONALES CONVOLUCIONALES

CNN aprendan patrones no lineales complejos, introduciendo factores de no linealidad que mejoran su capacidad de ajuste. Entre las funciones más comunes están el sigmoid, tanh, ReLU y sus variantes (*Leaky ReLU*, *Parametric ReLU*, *Randomized ReLU*, y *ELU*). En comparación con funciones como sigmoid y tanh, ReLU y sus variantes superan el problema de la desaparición del gradiente.

3. Capas de *Pooling*: Conocida también como capa de muestreo, incluye variantes como *max-pooling* donde se selecciona el valor máximo en una vecindad, *mean-pooling* toma el promedio de los valores y *pooling* estocástico selecciona aleatoriamente un valor dentro de la vecindad.
4. Capas Completamente Conectadas: La capa completamente conectada se encuentra al final de una CNN y convierte la información bidimensional en información unidimensional para la clasificación. Esta capa actúa de forma similar a las capas ocultas en los perceptrones multicapa (*Multilayer Perceptron*, MLP), generando la salida a través de combinaciones ponderadas de las neuronas previas.
5. Capa de Salida: Una vez que la clasificación final es lograda a través de la capa de salida de la CNN, las funciones de pérdida se utilizan para calcular el error predicho en los ejemplos de entrenamiento. Este error refleja la diferencia entre la salida real y la predicción. Las funciones de pérdida utilizan dos parámetros: la salida estimada por la CNN (predicción) y la salida real (etiqueta). Existen varios tipos de funciones de pérdida:
 - *Softmax Loss*: Usada en clasificación multiclase, mide el rendimiento generando probabilidades y utiliza softmax en la capa de salida.
 - *Euclidean Loss (MSE)*: Aplicada en problemas de regresión, mide el error cuadrático medio entre las predicciones y los valores reales.
 - *Hinge Loss*: Empleada en clasificación binaria, maximiza el margen entre clases, usada comúnmente en SVM [3, 36].
 - *Binary Cross-Entropy Loss (Entropía Cruzada Binaria)*: Se utiliza en tareas de clasificación binaria, en problemas donde la salida es una probabilidad, por ejemplo se usa para clasificar pares de imágenes como 'similares' o 'diferentes', comparando las probabilidades predichas con los resultados reales.
 - *Contrastive Loss (Pérdida Contrastiva)*: Se emplea en tareas de aprendizaje de similitudes, por ejemplo mide la similitud entre imágenes usando la distancia entre ellas. Esta función ajusta la

red para minimizar la distancia entre imágenes similares y maximizar la de imágenes diferentes.

- *Triplet Loss* (Pérdida de Tripletes): Es comúnmente utilizada en problemas de reconocimiento de patrones o identificación, utiliza un ancla, una imagen positiva y una negativa, ajustando la red para acercar el ancla a la imagen positiva y alejarla de la negativa, manteniendo un margen mínimo [7].

2.1.2. Fine-tuning

El *fine-tuning* ajusta modelos preentrenados para nuevos conjuntos de datos específicos, mejorando su rendimiento en tareas concretas. Se suelen congelar las primeras capas, que capturan características generales, mientras que las capas finales se ajustan para aprender las características del nuevo dominio [2]. Este proceso es un tipo de transferencia de conocimiento, útil cuando hay pocos datos etiquetados en el dominio específico, lo que permite una convergencia más rápida y una mayor precisión [32].

2.1.3. ResNet

ResNet, o red residual, es un tipo de arquitectura de aprendizaje profundo introducida por Kaiming He et al. en 2015, donde aborda el problema de la desaparición de gradientes en redes muy profundas utilizando conexiones de salto o conexiones residuales; estas conexiones permiten a la red aprender mapeos residuales en lugar de mapeos directos, lo que facilita el entrenamiento, como se observa en la Figura 1 donde se muestra el diseño de la arquitectura general. Además, en lugar de aprender una transformación completamente nueva para cada capa, las conexiones residuales permiten que el modelo aprenda la diferencia o 'residual' entre la entrada y la salida esperada, facilitando el aprendizaje [15].

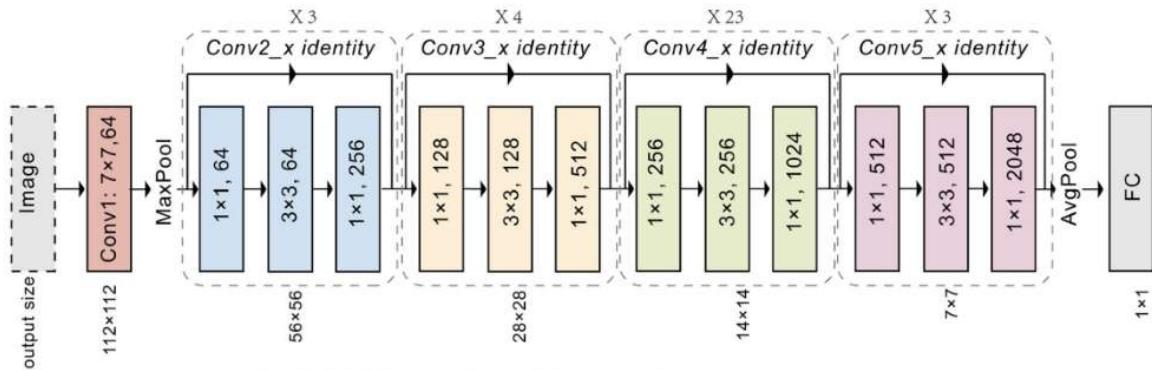


Figura 1: General Architecture Design [1]

2.2. Redes Siamesas

Las redes convolucionales siamesas fueron introducidas por Bromley y LeCun en los años 90, originalmente para resolver problemas de verificación de firmas como una tarea de correspondencia de imágenes. Estas redes constan de dos redes gemelas que aceptan entradas distintas, pero comparten parámetros, lo que asegura que imágenes similares no se mapen a ubicaciones muy diferentes en el espacio de características [8]. Posteriormente, LeCun et al. implementaron una función de energía contrastiva para reducir la energía de pares similares e incrementarla en los disímiles [24]. Alternativamente, enfoques como el de DeepFace utilizan la distancia L1 ponderada con una activación sigmoide y pérdida de entropía cruzada para entrenar la red [45].

Las redes convolucionales son particularmente efectivas en visión por computadora debido a su conectividad local, lo que reduce los parámetros y actúa como regularización implícita. Además, su capacidad para extraer características espaciales importantes ha sido clave en el éxito de tareas como el reconocimiento de imágenes [19, 21, 43].

2.2.1. Arquitectura de las redes siamesas

Las redes siamesas están formadas por dos subredes idénticas que procesan dos entradas diferentes. Cada subred utiliza una red neuronal convolucional (CNN) para extraer características de las imágenes y genera una representación comprimida a través de una capa totalmente conectada (FC). Ambas subredes comparten los mismos parámetros y capas de extracción.

Para medir la similitud entre las entradas, la red utiliza la distancia euclídea o la distancia Manhattan, estas funciones ayudan a identificar similitudes y diferencias entre las entradas, produciendo una codificación que se utiliza para comparar nuevas muestras. [19, 42].

2.3. Generación de Imágenes mediante Modelos Generativos

2.3.1. Modelos generativos de imágenes y su evolución

Los modelos generativos de imágenes son algoritmos de aprendizaje automático diseñados para crear contenido visual de forma autónoma al aprender patrones de datos visuales. Su desarrollo comenzó con redes generativas antagónicas (*Generative Adversarial Networks*, GANs) [14], que establecieron un enfoque basado en un proceso antagónico, donde emplean dos redes (un generador y un discriminador) en competencia para mejorar continuamente la calidad de las imágenes generadas.

2.3. GENERACIÓN DE IMÁGENES MEDIANTE MODELOS GENERATIVOS

Luego, el avance de los *autoencoders variacionales* (*Variational Auto-encoder*, VAEs) [18], proporcionó un marco probabilístico, permitiendo la generación de imágenes variadas y controlables mediante un espacio latente.

Posteriormente, se integra *StyleGAN* [17], que revolucionó el campo al producir imágenes de alta calidad y realismo, ampliando las posibilidades de personalización en la síntesis de imágenes.

Con la incorporación de mecanismos de atención y arquitecturas de transformadores como *Vision Transformer* (ViT) [11], los modelos ganaron capacidad para captar patrones de secuencias complejas, logrando representaciones más detalladas y contextualmente precisas.

Finalmente, los Modelos de Difusión de *Denoising* (*Denoising Diffusion Probabilistic Models*, DDPMs) [13, 16], introdujeron un método probabilístico para generar imágenes refinadas mediante un proceso iterativo de denoising. Estos modelos alcanzaron notoriedad en aplicaciones como *DALL·E* [35], que genera imágenes a partir de descripciones textuales, destacándose por su capacidad de control y su diversidad en las imágenes producidas [47].

2.3.2. Modelos Generativos de Difusión

Los Modelos Generativos de Difusión (*Diffusion Models*) surgieron como una nueva etapa en la generación de imágenes con la introducción de los *Denoising Diffusion Probabilistic Models* (DDPM) [16]. Funcionan agregando gradualmente ruido gaussiano a los datos originales en un proceso de difusión hacia adelante, y luego aprenden a eliminar ese ruido en un proceso de difusión inverso; a este proceso se le conoce como *denoising*.

En 2021, Song et al. optimizaron el proceso de *denoising* al trasladarlo al espacio latente, resultando en los *Denoising Diffusion Implicit Models* (DDIM) [44], lo cual mejoró significativamente la calidad generativa.

La evolución continuó con la integración de encoders avanzados, como CLIP (*Contrastive Language-Image Pre-Training*) [33], y métodos de condicionamiento como ILVR (*Iterative Latent Variable Refinement*) [10], que elevaron el rendimiento en tareas de generación de imágenes.

En 2022, empresas tecnológicas lanzaron frameworks basados en modelos de difusión, como *DALL·E-2* [34], Imagen [40] y *Stable Diffusion* [37], destinados a crear imágenes de alta calidad para fines comerciales.

El modelado de difusión de eliminación de ruido es un proceso de dos pasos:

1. Proceso de difusión hacia adelante: Este proceso es una cadena de Markov de pasos de difusión en la que se agrega lenta y aleatoriamente ruido a los datos originales.
2. Proceso de difusión inverso: Intenta revertir el proceso de difusión para generar datos originales a partir del ruido.

Proceso de difusión hacia adelante

En el proceso de difusión hacia adelante, como se muestra en la Figura 2, agregamos lentamente ruido gaussiano a la imagen de entrada X_0 a través de una serie de T pasos. Comenzamos muestreando un punto de datos X_0 de la distribución de datos reales $q(x)$ como $X_0 \sim q(x)$, y luego agregamos ruido gaussiano con una varianza β_t a X_{t-1} , produciendo una nueva variable latente X_t con distribución $q(X_t|X_{t-1})$ como se muestra en la Ecuación 2.1.

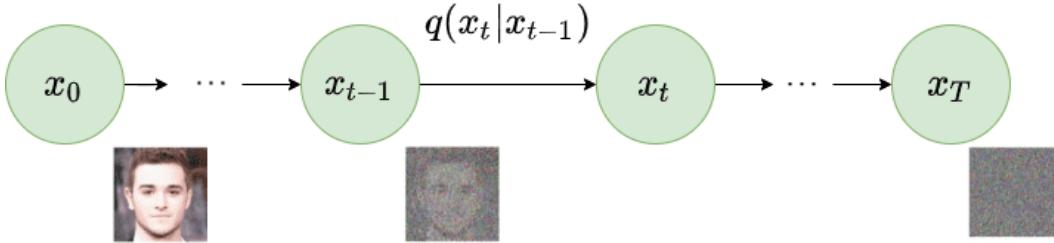


Figura 2: Proceso de difusión hacia adelante [16]

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1}) \quad (2.1)$$

$q(X_t|X_{t-1})$ se define por la media μ como en la Ecuación 2.2:

$$\mu_t = \sqrt{1 - \beta_t}X_{t-1} \quad (2.2)$$

y la matriz de covarianza Σ , donde $\sum_t = \beta_t I$ es la matriz identidad y Σ siempre será una matriz diagonal de varianzas. A medida que T se acerca a ∞ , X_t se convierte en una distribución gaussiana isotrópica.

Reparametrización

Aplicar $q(X_t|X_{t-1})$ y calcular X_t para un paso temporal arbitrario puede resultar muy costoso para un gran número de pasos. El truco de reparametrización soluciona este problema y nos permite muestrear X_t en cualquier paso temporal arbitrario de la siguiente distribución, como se muestra en la Ecuación 2.3:

$$X_t \sim q(x_t|x_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I}) \quad (2.3)$$

Proceso de difusión inverso

El proceso de difusión inverso, como se muestra en la Figura 3 consiste en recuperar los datos originales revirtiendo el proceso de ruido aplicado en la pasada hacia adelante. Estimar $q(X_t|X_{t-1})$ es difícil, ya que puede requerir

2.3. GENERACIÓN DE IMÁGENES MEDIANTE MODELOS GENERATIVOS

todo el conjunto de datos. Por eso, se puede usar un modelo p_θ (red neuronal) para aprender los parámetros. Para valores suficientemente pequeños de β_t , será gaussiano y puede obtenerse simplemente parametrizando la media y la varianza como se muestra en la Ecuación 2.4.

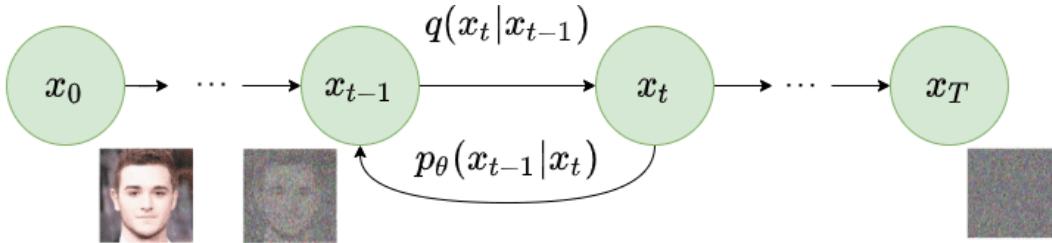


Figura 3: Proceso de difusión inverso [16]

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \quad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \quad (2.4)$$

Se entrena la red para predecir la media y la varianza en cada paso temporal. Aquí, $\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$ es la media, y $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)$ es la matriz de covarianza [27, 28, 48, 52].

2.3.3. *Stable Diffusion*

Stable Diffusion es un modelo de inteligencia artificial generativa, lanzado en 2022 [4], que permite crear imágenes fotorrealistas a partir de descripciones en texto. Su diseño se basa en un modelo de difusión en el espacio latente, lo que facilita la generación de imágenes de alta calidad sin requerir un poder computacional excesivo. Además de generar imágenes, este modelo se puede adaptar para crear videos y animaciones y permite personalización a través de aprendizaje por transferencia, lo cual permite obtener resultados específicos con solo algunas imágenes de referencia.

Arquitectura

La arquitectura de Stable Diffusion combina varios componentes clave que permiten transformar texto en imágenes. Estos son: el Codificador Automático Variacional, el proceso de Difusión Directa e Inversa, el Predictor de Ruido y el módulo de Acondicionamiento de Texto [4].

1. **Codificador Automático Variacional (VAE):** [18] en *Stable Diffusion* consiste en un codificador y un decodificador. Durante la generación, el codificador comprime una imagen de 512x512 píxeles en una representación latente de 64x64, lo que disminuye la carga de procesamiento al reducir la dimensionalidad. Este paso es crucial para que

2.3. GENERACIÓN DE IMÁGENES MEDIANTE MODELOS GENERATIVOS

el modelo pueda realizar la difusión en el espacio latente, un entorno más eficiente que el espacio de píxeles en cuanto a uso de memoria y potencia de procesamiento. Luego, el decodificador toma esta representación comprimida y la convierte de nuevo en una imagen de alta resolución, eliminando cualquier ruido residual.

2. Difusión Directa e Inversa:

- **Difusión Directa:** En este proceso, la imagen se degrada paulatinamente mediante la adición de ruido aleatorio hasta disolverse completamente. Esta fase se realiza durante el entrenamiento y permite que el modelo aprenda a gestionar distintos niveles de ruido en las imágenes. En el caso de *Stable Diffusion*, esta difusión directa no se utiliza en la fase de inferencia, a menos que se trate de una tarea de "imagen a imagen".
- **Difusión Inversa:** Esta etapa invierte el proceso de ruido de manera iterativa, eliminándolo gradualmente para reconstruir una imagen coherente. Gracias al entrenamiento con miles de millones de imágenes, el modelo puede generar contenido único a partir de cualquier descripción textual, lo que brinda gran flexibilidad para crear imágenes variadas y detalladas.

3. **Predictor de Ruido (U-Net):** El predictor de ruido es esencial para la eliminación del ruido en el espacio latente. *Stable Diffusion* utiliza una arquitectura U-Net [38], una red convolucional residual con conexiones de salto, que ayuda a restaurar detalles a partir de las representaciones ruidosas. Esta red, originalmente desarrollada para segmentación de imágenes biomédicas, en *Stable Diffusion* se adapta para predecir el nivel de ruido en cada paso iterativo. En cada iteración, la U-Net estima la cantidad de ruido en el espacio latente y lo elimina, ajustándose a las indicaciones textuales que guían el proceso.

4. **Acondicionamiento de Texto:** Para orientar la generación de imágenes, *Stable Diffusion* usa un módulo de acondicionamiento de texto basado en el modelo CLIP [33]. Este codificador convierte las descripciones textuales en embeddings, o representaciones vectoriales, que guían el proceso de eliminación de ruido en el U-Net. Las entradas textuales pueden contener hasta 75 tokens, y al configurar una semilla en el generador de números aleatorios, se pueden generar diferentes versiones de una imagen partiendo de una misma descripción.

5. **Scheduler:** *Stable Diffusion* cuenta con un *scheduler* que controla la adición y eliminación de ruido en las imágenes latentes. Este módulo asegura que el ruido se elimine de forma coherente y gradual en cada paso, lo que contribuye a obtener imágenes de alta calidad en cada iteración del proceso de difusión [25, 37, 53].

Stable Diffusion XL (SDXL)

El modelo *Stable Diffusion XL* (SDXL) es una mejora oficial sobre la versión 1.5, lanzado como software de código abierto. Este modelo representa un avance significativo en términos de tamaño y capacidad de generación de imágenes, con una arquitectura que incorpora dos modelos clave: el modelo base y el refinador.

El modelo base de SDXL cuenta con 3.5 mil millones de parámetros, mientras que el modelo combinado (base + refinador) alcanza 6.6 mil millones de parámetros, en comparación con los 0.98 mil millones de la versión 1.5. Este incremento en los parámetros implica un modelo considerablemente más robusto y capaz de generar imágenes de alta calidad y detalle.

La estructura de SDXL sigue una secuencia en la que primero actúa el modelo base, encargado de establecer la composición global de la imagen, y luego el refinador, que añade los detalles más finos como se muestra en la Figura 4. Esta secuencia es opcional, por lo que se puede usar el modelo base de manera independiente si así se desea.

Además, SDXL incluye una combinación avanzada de modelos de lenguaje, que integra el modelo *OpenClip* más grande (ViT-G/14) y el CLIP ViT-L de OpenAI. Este enfoque facilita el uso de *prompts* y mejora la capacidad de respuesta a las indicaciones del usuario, superando la dificultad de ajuste de *prompts* observada en la versión 2.0, que emplea únicamente *OpenClip*.

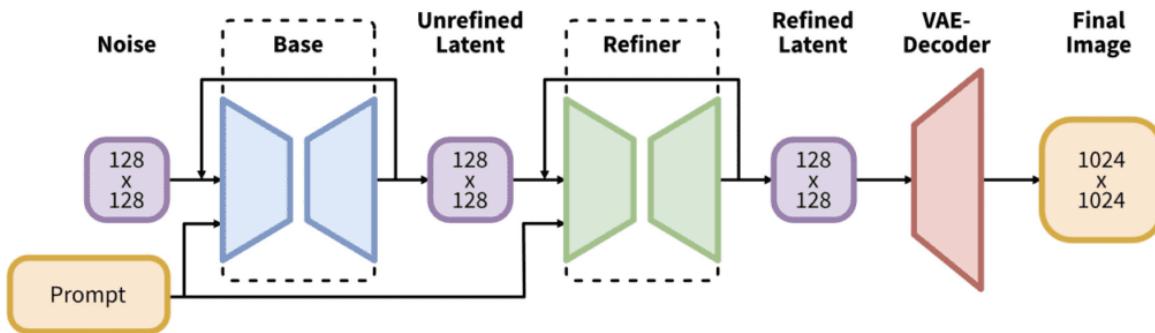


Figura 4: Arquitectura del modelo base y del modelo refinador. [31]

Otra mejora importante es la nueva técnica de acondicionamiento de tamaño de imagen de SDXL, que permite aprovechar imágenes de entrenamiento de menos de 256×256 píxeles, aumentando el conjunto de datos útil al no descartar un 39 % de las imágenes.

Asimismo, la arquitectura de U-Net, que es la base del modelo de difusión, ahora es tres veces más grande, permitiendo una mayor precisión en la generación de imágenes. El tamaño de imagen predeterminado de SDXL es de 1024×1024 píxeles, cuadruplicando el tamaño de la versión 1.5, que se limitaba a 512×512 píxeles.

2.3. GENERACIÓN DE IMÁGENES MEDIANTE MODELOS GENERATIVOS

En general, los usuarios han mostrado una clara preferencia por el modelo SDXL en comparación con el modelo base 1.5 y 2.1, según un estudio de Stability AI como se muestra en la Figura 5 [6, 31].

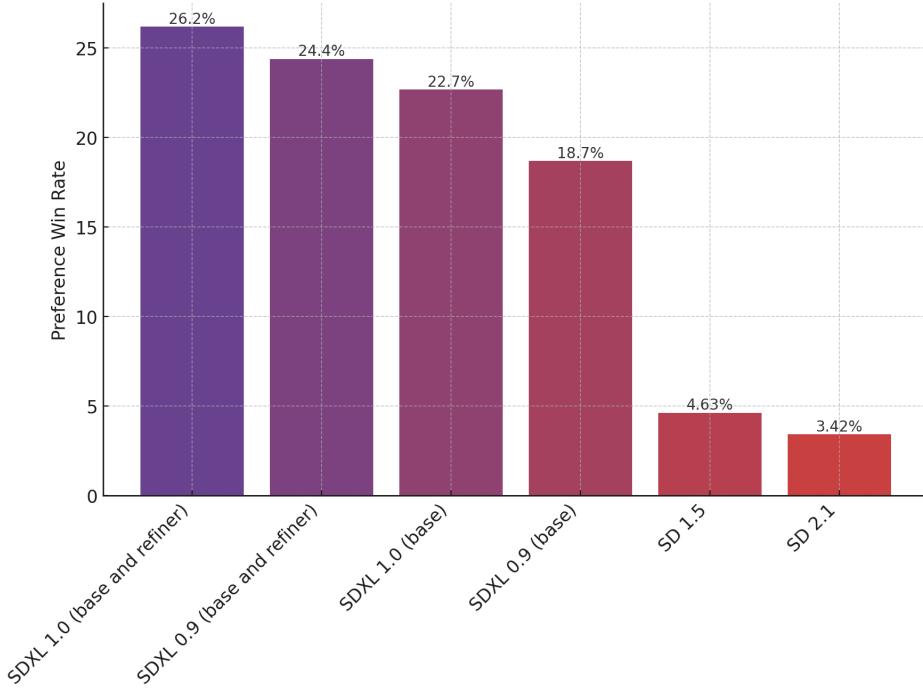


Figura 5: Comparación de las preferencias de los usuarios entre SDXL y *Stable Diffusion* 1.5 y 2.1 [31]

Para optimizar la generación de imágenes con el modelo *Stable Diffusion XL* (SDXL), se presentan algunas recomendaciones para la construcción de *prompts*:

- **Descripciones detalladas en lenguaje natural:** A diferencia del modelo v1, que interpretaba los prompts como un conjunto de palabras clave, el modelo SDXL incorpora un modelo de lenguaje avanzado que permite una interpretación más precisa de descripciones en lenguaje natural. Es recomendable especificar la imagen con el mayor detalle posible, utilizando oraciones completas, aunque el uso de palabras clave sigue siendo efectivo.
- **Uso moderado de prompts negativos:** A diferencia de los modelos v1, que a menudo requerían prompts negativos extensos, el modelo SDXL minimiza esta necesidad. Se sugiere incluir solo los elementos específicos que se desean evitar, como términos relacionados con estilos no deseados (por ejemplo, 'dibujo animado' en imágenes de estilo fotográfico) o cualquier objeto que se prefiera excluir de la generación.

- **Ajuste cuidadoso de los pesos de palabras clave:** El modelo SDXL es más sensible a los ajustes en los pesos de las palabras clave. Al asignar mayor o menor importancia a ciertos términos, por ejemplo, mediante el formato (palabra clave: 1.1), para aumentar el énfasis en un 10 %, se recomienda limitar el ajuste; valores superiores a 1.4 suelen ser innecesarios y podrían afectar la precisión de la generación [5].

2.4. Descriptores de imágenes

Los descriptores de imágenes son herramientas que permiten representar, de manera cuantificable, características visuales específicas de una imagen, como el color, la textura o la forma. Estos descriptores estandarizados se utilizan en aplicaciones de búsqueda y clasificación, ya que permiten la comparación de contenido visual en grandes bases de datos de imágenes. En el estándar MPEG-7, los descriptores de imágenes están diseñados para ofrecer una representación compacta y precisa, lo que optimiza la interoperabilidad entre diferentes aplicaciones multimedia y facilita el procesamiento y la recuperación de información visual. Los descriptores proporcionan una 'firma' de cada imagen, lo cual es útil para reconocer y categorizar contenido visual en función de sus atributos específicos, incluso bajo diversas condiciones de iluminación o tamaño [39].

2.4.1. Tipos de descriptores

Los tipos de descriptores de imágenes se pueden clasificar en función de los atributos visuales que representan, destacando los descriptores de color, textura y forma o bordes. Cada tipo de descriptor captura información única que es fundamental para analizar diferentes aspectos de la imagen [39]. Por ejemplo, los descriptores de color se utilizan para capturar la distribución cromática en una imagen, lo que permite identificar patrones de color específicos; los descriptores de textura se utilizan para representar patrones repetitivos y variaciones estructurales en la imagen, y los descriptores de forma y bordes, ya sea basados en el contorno o en la región, se utilizan en la disposición y contorno de los objetos dentro de la imagen, de modo que identifican formas y estructuras.

Descriptores de color

Entre los descriptores de color más importantes en el estándar MPEG-7 se encuentran el Histograma de Color Escalable (SCD) y el Histograma de Estructura de Color (CSD), que se utilizan en los espacios de color HSV y HMMD, respectivamente. El SCD permite una representación de la distribución de color y utiliza transformadas como la de Haar para mejorar la precisión. El CSD, por su parte, ayuda a identificar patrones de color en áreas

2.4. DESCRIPTORES DE IMÁGENES

localizadas, para segmentar o categorizar partes específicas de la imagen según su color. Otros descriptores incluyen el Descriptor de Color Dominante (DCD), que identifica los colores predominantes, y el Descriptor de Disposición de Color (CLD), que organiza los colores en una cuadrícula para capturar su distribución espacial. [39]

Estos descriptores, además de otros como HueSIFT y OpponentSIFT, se utilizan principalmente en el reconocimiento de escenas y objetos, ya que permiten comparar imágenes de manera precisa incluso bajo condiciones de iluminación variables, siendo descriptores robustos. [46]

Descriptores de textura

Los descriptores de textura, por su parte, permiten caracterizar aspectos como la regularidad, la finura y la direccionalidad en una imagen. Entre ellos se encuentra el Descriptor de Textura Homogénea (HTD), que analiza frecuencias espaciales en el dominio de Fourier, para identificar y comparar texturas homogéneas; el Descriptor de Navegación de Textura, que categoriza la textura de acuerdo con parámetros como la dirección y el grosor, entre otros. Los descriptores de textura ayudan a identificar características dentro de una imagen que no se capturan con precisión mediante descriptores de color o de bordes.

Descriptores de bordes y formas

Los descriptores de forma y bordes permiten analizar la estructura y el contorno de los objetos dentro de la imagen, clasificándose en métodos basados en el contorno y en la región. Los descriptores basados en contorno, como el código de cadena y los descriptores de Fourier, se centran en la representación de los límites de los objetos en una imagen y proporcionan una representación matemática robusta que es invarianta a cambios de rotación y escalado.

Por otro lado, los descriptores basados en región, como los momentos invariantes y los momentos de Zernike, permiten caracterizar tanto el contorno como el interior de la forma, asegurando la invariancia geométrica y facilitando la comparación de imágenes en diferentes orientaciones y tamaños.

Descriptores semánticos

Los descriptores semánticos van más allá de los atributos visuales básicos como el color, la textura y la forma para identificar y capturar el contenido significativo de una imagen. Estos descriptores suelen utilizar técnicas de aprendizaje profundo, como redes neuronales convolucionales (CNN), para extraer representaciones abstractas de características visuales de bajo nivel y asignarlas a conceptos semánticos como categorías de objetos,

escenas o emociones en la imagen. Por ejemplo, los modelos previamente entrenados como VGG, ResNet o Inception se utilizan ampliamente en tareas de clasificación de imágenes para capturar características semánticas que representan patrones complejos y contextos específicos[26]. Estos descriptores son particularmente útiles en aplicaciones de recuperación de imágenes y análisis de contenido, ya que permiten comparar imágenes a nivel de significado en lugar de solo similitud visual.

A diferencia de los descriptores tradicionales que solo representan aspectos visuales, los descriptores semánticos tienen la ventaja de estructuras de red profundas y jerárquicas para extraer características de alto nivel que distinguen categorías complejas en imágenes. Actualmente, se están desarrollando arquitecturas de redes neuronales más avanzadas que integran información contextual y multimodal, ampliando aún más la gama de descriptores semánticos [22].

2.5. Estado del Arte

A medida que ha crecido el interés en la inteligencia artificial generativa y su capacidad para crear imágenes, se han planteado desafíos en términos de derechos de autor, especialmente al imitar estilos artísticos protegidos.

Modelos como *Stable Diffusion* han sido objeto de estudio, ya que generan imágenes que a menudo capturan estilos distintivos de artistas reconocidos, ha demostrado una capacidad para replicar estilos artísticos con un 81 % de precisión y alcanzando un 90 % de similitud con obras originales [9]. Esta capacidad de imitación ha sido estudiada en investigaciones más profundas, como en [54], que reveló que el 70 % de las imágenes generadas por modelos de difusión contenían similitudes significativas con contenido protegido por derechos de autor.

Esto ha llevado a investigadores a replantearse el concepto de 'infracción artística', abordándolo no como una simple coincidencia de imágenes, sino como un problema de clasificación de estilos [30].

En respuesta a estos desafíos, surgieron herramientas especializadas como *StyleAuditor* [12], que alcanzó una precisión superior al 90 % en la identificación de imitaciones artísticas utilizando redes neuronales preentrenadas como VGG.

Los desarrollos más recientes, como *ArtSavant* [30], han integrado un enfoque combinando *DeepMatch* y *TagMatch* basados en CLIP para lograr un 89.3 % de precisión en la clasificación de autoría y un 82.5 % en la identificación de estilos artísticos. La Tabla 1 presenta algunos trabajos relevantes para contextualizar este panorama:

Tabla 1: Trabajo relacionado con similitud de imágenes de IAs

Referencia	Descripción	Resultados Clave
<i>ArtSavant</i> [30]	Se usan modelos como <i>DeepMatch</i> y <i>TagMatch</i> con el dataset WikiArt para evaluar si imágenes generadas por IA replican estilos de 372 artistas, analizando posibles infracciones de derechos de autor.	<i>DeepMatch</i> clasifica autorías con 89.3 % de precisión y <i>TagMatch</i> identifica estilos con 61.6 % top-1 y 82.5 % top-5. Se propone analizar infracciones detectando la replicación de estilos con <i>Stable Diffusion</i> como generador.
<i>StyleAuditor</i> [12]	Se utilizan redes preentrenadas como VGG y el modelo generativo <i>Stable Diffusion</i> para auditar la imitación del estilo artístico de 30 artistas, comparando sus obras públicas con imágenes generadas.	Con una precisión superior al 90% y una baja tasa de falsos positivos, la herramienta resulta eficiente para auditar imitaciones de estilo artístico.
<i>Measuring the Success of Diffusion Models at Imitating Human Artists</i> [9]	Se utiliza <i>Stable Diffusion</i> y CLIP (zero-shot) para evaluar la capacidad de modelos de difusión de imitar estilos de 70 artistas profesionales, analizando implicaciones legales mediante imágenes generadas con prompts específicos.	<i>Stable Diffusion</i> logra una precisión del 81 % al imitar estilos artísticos y un 90 % de similitud entre imitaciones y obras originales.
<i>On Copyright Risks of Text-to-Image Diffusion Models</i> [54]	Se utilizan <i>Stable Diffusion</i> , CLIP y GPT-3.5 para evaluar la infracción de derechos de autor en modelos de difusión de <i>text to image</i> , analizando imágenes generadas con prompts genéricos (películas, videojuegos, logotipos) y comparándolas con imágenes con derechos de autor anotadas manualmente.	El 70 % de las imágenes generadas contenían similitudes significativas con contenido con derechos de autor. Los <i>prompts</i> generados no tenían alta sensibilidad semántica respecto a los temas objetivo.

El uso de redes neuronales siamesas para la comparación de imágenes ha demostrado su efectividad con precisiones de hasta 94 % *mean Average Precision* (mAP) en aplicaciones como la recuperación y el reconocimiento de patrones. Estas redes se han utilizado gracias a su capacidad de generalización a conjuntos de datos no vistos durante el entrenamiento, por lo que resulta un modelo robusto para la evaluación y comparación de características visuales en imágenes de obras de arte, incluyendo la detección de similitudes con contenido protegido.

Un caso destacado es el uso de redes siamesas para mejorar la coincidencia de imágenes en grandes bases de datos, como se observa en el trabajo [29], donde se demuestra que las redes siamesas con pérdida contrastiva *HybridCNN* y *sHybridCNN*, ofrecen un mejor rendimiento que otros enfoques como AlexNet, incluso cuando las etiquetas de las imágenes son imperfectas. De manera similar, se ha utilizado este tipo de redes para mejorar la recuperación de imágenes y la localización de patrones en colecciones de documentos, como se observa en el uso de una red siamesa entrenada en el conjunto de datos Tobacco800, que obtuvo un mAP de 0.94 y una precisión de 0.83 en la localización de patrones. La Tabla ?? presenta algunos trabajos relacionados a la comparación de imágenes utilizando Redes Siamese:

Tabla 2: Trabajo relacionado con similitud de imágenes de IAs

Referencia	Descripción	Resultados Clave
"Deep Siamese Network with Multi-level Similarity Perception for Person Re-identification" [41]	Se utiliza una red neuronal profunda tipo Siamese con similitud a múltiples niveles para mejorar la re-identificación de personas, utilizando datos de CUHK03, Market-1501 y CUHK01 con imágenes de varias cámaras.	Los resultados obtenidos fueron: 85.7 % en rank-1 para CUHK03, 83.6 % en rank-1 para CUHK03, y 81.9 % en rank-1 y 63.6 % en mAP para Market-1501.
Recuperación de imágenes y localización de patrones usando Redes Neuronales Siamese (SNN) [49]	Se utiliza una Red Siamese con aprendizaje por transferencia para mejorar la recuperación de imágenes en el conjunto Tobaccos00, que contiene 1,290 imágenes de documentos y 418 consultas de búsqueda.	Los resultados muestran una precisión promedio (mAP) de 0.94 en recuperación, 0.83 en localización de patrones, un IoU de 0.7, y superan a los métodos tradicionales en recuperación y localización de patrones en imágenes de documentos.
Siamese Network Features for Image Matching [29]	Se usan redes neuronales Siamese, Hybrid-CNN y sHybridCNN para mejorar la coincidencia de imágenes en grandes bases de datos con pares etiquetados como coincidentes o no.	sHybridCNN supera a HybridCNN y AlexNet en AUC, mejorando la coincidencia de imágenes con etiquetas imperfectas y mostrando buen rendimiento en datos no vistos.

CAPÍTULO 3

METODOLOGÍA

La presente investigación sigue una metodología estructurada en siete fases principales, para implementar y evaluar un modelo de comparación de similitud entre obras de arte originales y sus contrapartes generadas por IA. La Figura 6 muestra el diagrama de metodología con las siete fases del proyecto.

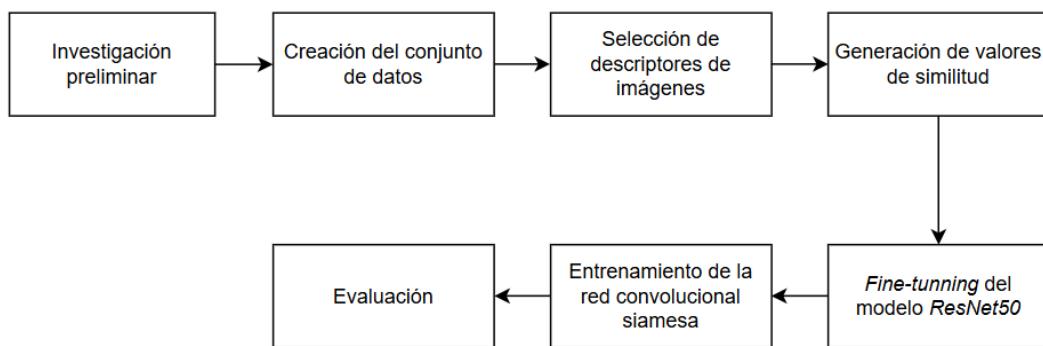


Figura 6: Diagrama de metodología del proyecto. Fuente propia.

En sus fases iniciales este estudio es exploratorio, dado que busca establecer relaciones entre las características visuales de las imágenes originales y generadas. A continuación, se describe detalladamente cada fase del proceso.

3.1. Investigación preliminar.

La investigación comienza con la revisión de la literatura científica, centrándose en tres áreas: redes convolucionales siamesas, métodos actuales de comparación de imágenes, y proyectos relacionados con la similitud entre imágenes generadas por IA y obras de arte originales; con el fin de obtener los fundamentos teóricos y el estado del arte para guiar el proyecto.

3.2. Creación del conjunto de datos.

En esta fase del proyecto se realiza un proceso sistemático para la generación del conjunto de datos experimental. Este proceso involucra la recopilación de obras de arte originales desde WikiArt, seleccionando piezas representativas de diferentes períodos y estilos artísticos, como la pintura de Nicholas-roerich, mostrada en la Figura 7, disponible en dicho dataset.



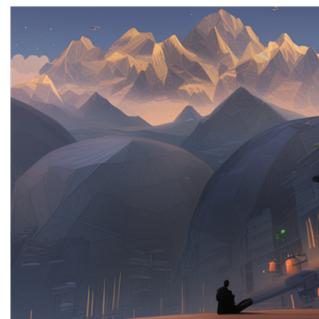
Figura 7: Pintura de Nicholas-roerich. Wikiart.

Posteriormente, se genera una serie de imágenes utilizando *Stable Diffusion XL Refiner 1.0*, basándose en *prompts* derivados de las descripciones originales de las obras de arte. Tal como se muestra en la Figura 8, se introducen los siguientes prompts:

- "Add mythical creatures such as dragons and fairies to the scene" (.^ñade criaturas míticas, como dragones y hadas, a la escena").
- "Reimagine in a futuristic setting, with advanced technology and neon architecture"(Reimagina en un entorno futurista, con tecnología avanzada y arquitectura neón").



"Add mythical creatures such as dragons and fairies to the scene."
Stable Diffusion



"Reimagine in a futuristic setting, with advanced technology and neon architecture." Stable Diffusion.

Figura 8: Imágenes generadas con referencia de la pintura de Nicholas-roerich. *Stable Diffusion*.

3.3. Selección de descriptores de imágenes.

En esta etapa se analizan diversos descriptores de características visuales, como se muestran en la Figura 9, tales como descriptores de color, enfocados en las paletas dominantes y la distribución cromática; los de textura, que evalúan patrones y granularidad; los de forma, centrados en contornos y geometrías; los basados en bordes, que analizan la definición y continuidad; y los semánticos, que consideran elementos compositivos. La selección final de estos descriptores se realizará en función de su eficacia para cuantificar la similitud entre las imágenes originales y las generadas.



Figura 9: Proceso de selección de descriptores de imágenes. Fuente propia.

3.4. Generación de valores de similitud.

A partir de los descriptores seleccionados, se calculan valores de similitud mediante la aplicación de factores ponderados a cada descriptor, asignados según su importancia relativa en la evaluación visual. Los resultados obtenidos permitirán generar un valor numérico que será la similitud global entre cada par de imágenes, tal como se muestra en la Figura 10.

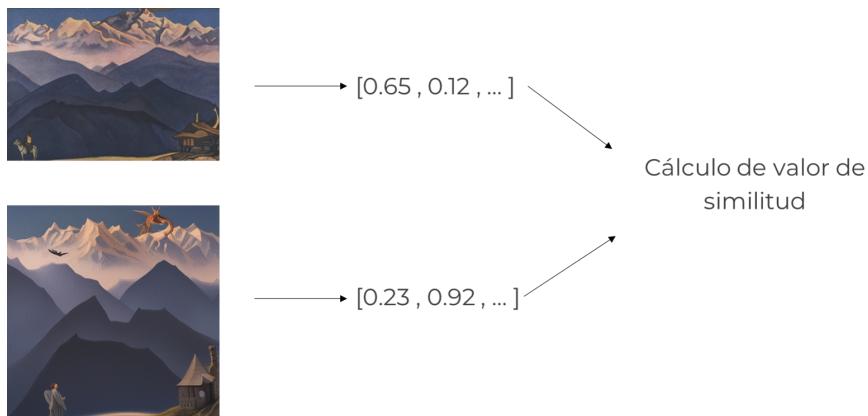


Figura 10: Proceso de generación de valores de similitud. Fuente propia.

3.5. **Fine-tuning del modelo ResNet50.**

El ajuste fino del modelo *ResNet50* se lleva a cabo utilizando una versión preentrenada de la arquitectura, y se especializará para la clasificación de géneros artísticos mediante el uso de imágenes del conjunto de datos WikiArt. Para ello se divide el conjunto de datos en tres subconjuntos: entrenamiento, validación y prueba.

Después, se optimizan hiperparámetros, como la tasa de aprendizaje, el tamaño de lote y el número de épocas, para maximizar la precisión del modelo. Una vez completado el entrenamiento, los pesos ajustados del modelo se guardan para utilizarlos en la red siamesa.

3.6. **Implementación y entrenamiento de la Red Siamesa.**

Durante esta fase se implementa la red convolucional siamesa empleando los pesos adaptados del modelo *ResNet50* en el *fine-tuning*. El modelo estará configurado para generar un valor de regresión que cuantifique la similitud entre pares de imágenes, y será la métrica para evaluar la correspondencia visual entre ambas.

3.7. **Evaluación.**

En el estado del arte, se utilizan diversas métricas estándar para evaluar modelos de regresión, tales como:

- Error Cuadrático Medio

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3.1)$$

- Error Absoluto Medio

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|. \quad (3.2)$$

La evaluación del modelo implementa alguna de estas métricas para analizar su rendimiento. Además, se lleva a cabo un análisis comparativo entre los resultados obtenidos por la red siamesa y los valores de similitud preliminares basados en descriptores. Este proceso de evaluación permite obtener una medida objetiva del desempeño del sistema.

REFERENCIAS

- [1] Allan, K. (2023). Understanding ResNet: A Milestone in Deep Learning and Image Recognition. <https://www.ikomia.ai/blog/mastering-resnet-deep-learning-image-recognition>
- [2] Alshalali, T., & Josyula, D. (2018). Fine-Tuning of Pre-Trained Deep Learning Models with Extreme Learning Machine. *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, 469-473. <https://doi.org/10.1109/CSCI46756.2018.00096>
- [3] Alzubaidi, L., Zhang, J., Humaidi & et al., A. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(53), 2196-1115. <https://doi.org/10.1186/s40537-021-00444-8>
- [4] Amazon. (s.f.). What is Stable Diffusion? - Stable Diffusion AI Explained - AWS — aws.amazon.com. <https://aws.amazon.com/en/what-is/stable-diffusion/>
- [5] Andrew. (2024a). 15 Stable Diffusion XL prompts + tips. <https://stable-diffusion-art.com/sdxl-prompts/>
- [6] Andrew. (2024b). Stable Diffusion XL 1.0 model. <https://stable-diffusion-art.com/sdxl-model/>
- [7] Anurag. (2023, noviembre). Understanding Siamese Networks: A Comprehensive Introduction. <https://www.analyticsvidhya.com/blog/2023/08/introduction-and-implementation-of-siamese-networks/>
- [8] Bromley, J., Bentz, J. W., Bottou, L., Guyon, I. M., LeCun, Y., Moore, C., Säckinger, E., & Shah, R. (1993). Signature Verification Using A "Siamese"Time Delay Neural Network. *Int. J. Pattern Recognit. Artif. Intell.*, 7, 669-688. <https://api.semanticscholar.org/CorpusID:16394033>
- [9] Casper, S., Guo, Z., Mogulothu, S., Marinov, Z., Deshpande, C., Yew, R. J., Dai, Z., & Hadfield-Menell, D. (2023). Measuring the Success of Diffusion Models at Imitating Human Artists. *International Conference on Machine Learning*.
- [10] Choi, J., Kim, S., Jeong, Y., Gwon, Y., & Yoon, S. (2021). ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models. *arXiv e-prints*, Artículo arXiv:2108.02938, arXiv:2108.02938. <https://doi.org/10.48550/arXiv.2108.02938>

REFERENCIAS

- [11] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv, abs/2010.11929*. <https://api.semanticscholar.org/CorpusID:225039882>
- [12] Du, L., Zhu, Z., Chen, M., Ji, S., Cheng, P., Chen, J., & Zhang, Z. (2024). WIP: Auditing Artist Style Pirate in Text-to-image Generation Models. *CISPA Helmholtz Center for Information Security, Saarbrucken*.
- [13] Geng, D., Park, I., & Owens, A. (2024). Visual anagrams: Generating multi-view optical illusions with diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24154-24163.
- [14] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.
- [15] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778. <https://doi.org/10.1109/CVPR.2016.90>
- [16] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. *arXive-prints*, Artículo arXiv:2006.11239, arXiv:2006.11239. <https://doi.org/10.48550/arXiv.2006.11239>
- [17] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2019). Analyzing and Improving the Image Quality of StyleGAN. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8107-8116. <https://api.semanticscholar.org/CorpusID:209202273>
- [18] Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. *CoRR, abs/1312.6114*. <https://api.semanticscholar.org/CorpusID:216078090>
- [19] Koch, G. R. (2015). Siamese Neural Networks for One-Shot Image Recognition. <https://api.semanticscholar.org/CorpusID:13874643>
- [20] Krichen, M. (2023). Convolutional Neural Networks: A Survey. *Computers*, 12(8). <https://doi.org/10.3390/computers12080151>
- [21] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017a). ImageNet classification with deep convolutional neural networks. *Commun. ACM*, 60(6), 84-90. <https://doi.org/10.1145/3065386>
- [22] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017b). ImageNet classification with deep convolutional neural networks. *Commun. ACM*, 60(6), 84-90. <https://doi.org/10.1145/3065386>
- [23] Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324. <https://doi.org/10.1109/5.726791>
- [24] LeCun, Y., Chopra, S., Hadsell, R., Ranzato, A., & Huang, F. J. (2006). A Tutorial on Energy-Based Learning. <https://api.semanticscholar.org/CorpusID:8531544>

- [25] Lee, S., Hoover, B., Strobelt, H., Wang, Z. J., Peng, S., Wright, A., Li, K., Park, H., Yang, H., & Chau, D. H. (2023). Diffusion explainer: Visual explanation for text-to-image stable diffusion. *arXiv preprint arXiv:2305.03509*.
- [26] Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., & Pietikäinen, M. (2018). Deep Learning for Generic Object Detection: A Survey. *arXiv e-prints*, Artículo arXiv:1809.02165, arXiv:1809.02165. <https://doi.org/10.48550/arXiv.1809.02165>
- [27] Ma, Z., Zhang, Y., Jia, G., Zhao, L., Ma, Y., Ma, M., Liu, G., Zhang, K., Li, J., & Zhou, B. (2024). Efficient Diffusion Models: A Comprehensive Survey from Principles to Practices. <https://api.semanticscholar.org/CorpusID:273351124>
- [28] Manjurul Ahsan, M., Raman, S., Liu, Y., & Siddique, Z. (2024). A Comprehensive Survey on Diffusion Models and Their Applications. *arXiv e-prints*, Artículo arXiv:2408.10207, arXiv:2408.10207. <https://doi.org/10.48550/arXiv.2408.10207>
- [29] Melekhov, I., Kannala, J., & Rahtu, E. (2017). Siamese Network Features for Image Matching. *IEEE*.
- [30] Moayeri, M., Basu, S., Balasubramanian, S., Kattakinda, P., Chengini, A., Brauneis, R., & Feizi, S. (2024). Rethinking Artistic Copyright Infringements in the Era of Text-to-Image Generative Models. *University of Maryland, Computer Science Department*.
- [31] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., & Rombach, R. (2023). Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.
- [32] Radenović, F., Tolias, G., & Chum, O. (2019). Fine-Tuning CNN Image Retrieval with No Human Annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7), 1655-1668. <https://doi.org/10.1109/TPAMI.2018.2846566>
- [33] 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. *International Conference on Machine Learning*. <https://api.semanticscholar.org/CorpusID:231591445>
- [34] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv e-prints*, Artículo arXiv:2204.06125, arXiv:2204.06125. <https://doi.org/10.48550/arXiv.2204.06125>
- [35] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-Shot Text-to-Image Generation. *ArXiv*, abs/2102.12092. <https://api.semanticscholar.org/CorpusID:232035663>
- [36] A review of convolutional neural network architectures and their optimizations. (2022). *Artificial Intelligence Review*, 56, 1-65. <https://doi.org/10.1007/s10462-022-10213-5>

REFERENCIAS

- [37] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2021). High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv e-prints*, Artículo arXiv:2112.10752, arXiv:2112.10752. <https://doi.org/10.48550/arXiv.2112.10752>
- [38] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv e-prints*, Artículo arXiv:1505.04597, arXiv:1505.04597. <https://doi.org/10.48550/arXiv.1505.04597>
- [39] S., M. B., Ohm, J.-R., Vasudevan, V. V., & Yamada, A. (2001). Color and Texture Descriptors. *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY*.
- [40] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghase-mipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M. (2022). Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *ArXiv, abs/2205.11487*. <https://api.semanticscholar.org/CorpusID:248986576>
- [41] Shen, C., Jin, Z., Zhao, Y., Fu, Z., Jiang, R., Chen, Y., & Hua, X. (2017). Deep Siamese Network with Multi-level Similarity Perception for Person Re-identification. *IEEE*.
- [42] Sherly A P, P. A. R. (2024). Siamese Augmented Network (SAuGNet) for JPEG Steganalysis. <https://doi.org/10.36227/techrxiv.171994747.74344541/v1>
- [43] Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. <https://arxiv.org/abs/1409.1556>
- [44] Song, J., Meng, C., & Ermon, S. (2020). Denoising Diffusion Implicit Models. *ArXiv, abs/2010.02502*. <https://api.semanticscholar.org/CorpusID:222140788>
- [45] Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the Gap to Human-Level Performance in Face Verification. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1701-1708. <https://doi.org/10.1109/CVPR.2014.220>
- [46] van de Sande, K. E. A., Gevers, T., & Snoek, C. G. M. (2009). Evaluating Color Descriptors for Object and Scene Recognition. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*.
- [47] Wang, B., Chen, Q., & Wang, Z. (2024). Diffusion-Based Visual Art Creation: A Survey and New Perspectives. *ArXiv, abs/2408.12128*. <https://api.semanticscholar.org/CorpusID:271924425>
- [48] Weng, L. (2021). What are diffusion models? *lilianweng.github.io*. <https://lilianweng.github.io/posts/2021-07-11-diffusion-models/>
- [49] Wiggers, K. L., Britto Jr., A. S., Heutte, L., Koerich, A. L., & Oliveira, L. S. (2019). Image Retrieval and Pattern Spotting using Siamese Neural Network. *ArXiv*.
- [50] WikiArt. (s.f.). WikiArt [[Online; accessed 13-September-2024]].

- [51] Wikipedia contributors. (2024). Stable Diffusion—Wikipedia, The Free Encyclopedia [[Online; accessed 13-September-2024]].
- [52] Xing, Z., Feng, Q., Chen, H., Dai, Q., Hu, H., Xu, H., Wu, Z., & Jiang, Y.-G. (2024). A Survey on Video Diffusion Models [Just Accepted]. *ACM Comput. Surv.* <https://doi.org/10.1145/3696415>
- [53] Zhang, L., Rao, A., & Agrawala, M. (2023). Adding Conditional Control to Text-to-Image Diffusion Models. *arXive-prints*, Artículo arXiv:2302.05543, arXiv:2302.05543. <https://doi.org/10.48550/arXiv.2302.05543>
- [54] Zhang, Y., Teoh, T. T., Lim, W. H., Wang, H., & Kawaguchi, K. (2024). On Copyright Risks of Text-to-Image Diffusion Models. *National University of Singapore*.

REFERENCIAS

COLABORADORES

Directores

Asesor Dr. Ricardo Ramos Aguilar
Profesor en la carrera de Ingeniería en Inteligencia Artificial en
UPIIT-IPN.
email: rramosa@ipn.mx



Firma: _____

Asesor Dr. Jesús García Ramírez
Profesor en la carrera de Ingeniería en Inteligencia Artificial en
UPIIT-IPN.
email: jegarciara@ipn.mx



Firma: _____

Alumnos

Diego Castro Elvira
Alumno de la carrera de Ingeniería en Inteligencia Artificial en el
Instituto Politécnico Nacional.
Boleta: 2022710168
Tel. 2411081478
email: diego.castro.elvira@gmail.com



Firma: _____

Navil Pineda Rugerio
Alumna de la carrera de Ingeniería en Inteligencia Artificial en el
Instituto Politécnico Nacional.
Boleta: 2022710240
Tel. 2461509006
email: naviladenip@gmail.com



Firma: _____