Departamento de Posgrado en Ingeniería e Innovación

CETYS UNIVERSIDAD

Título de Tesis

COLORES QUE COBRAN VIDA: APLICACIÓN DE REDES NEURONALES PROFUNDAS EN LA COLORACIÓN DE IMÁGENES

Que para obtener el grado de

MAESTRÍA EN INGENIERÍA E INNOVACIÓN

Presenta:

Norma Vanessa Argueta Vázquez

Director de Tesis:
Dr. Ulises Orozco Rosas

Co-Director de tesis:
Dra. Kenia Picos Espinoza

Tijuana, Baja California a 17 de Junio de 2025.

ACTA DE REVISIÓN DE TESIS

En la ciudad de Tijuana, Baja California siendo el día 17 de junio del 2025 se reunieron los miembros del Comité de Revisión de Tesis designada en el Departamento de Posgrado de CETYS Universidad Campus Tijuana para examinar la tesis titulada:

COLORES QUE COBRAN VIDA: APLICACIÓN DE REDES NEURONALES PROFUNDAS EN LA COLORACIÓN DE IMÁGENES

Tesis para obtener el grado académico de la Maestría en Ingeniería e Innovación, presentada por la alumna:

Norma Vanessa Argueta Vázquez

Los miembros del Comité de Revisión de Tesis manifestaron **APROBAR LA TESIS**, en virtud que satisface con los requisitos establecidos por el Reglamento de Posgrado.

COMITÉ DE TESIS

Dr. Ulises Orozco Rosas
Director de Tesis

Dra. Kenia Picos Espinoza
Co-directora de Tesis

Dr. Adán Hirales Carbajal
Miembro del Comité

Dr. Ricardo Martínez Soto
Coordinador del Posgrado en Ingeniería
Miembro del Comité

DEDICATORIA

A mi mamá,
Gracias por creer en mí incluso cuando yo misma dudaba. Por tu amor incondicional, por tu apoyo constante, y por ser mi fuerza silenciosa en cada paso de este camino. Este logro también es tuyo, porque sin ti, este viaje de crecimiento personal no habría sido posible.
Con todo mi amor y gratitud.

AGRADECIMIENTOS

Al Dr. Ulises Orozco Rosas,
Expreso mi más sincero agradecimiento por su valioso acompañamiento a lo largo de este proyecto. Su apoyo constante, guía experta y paciencia inquebrantable fueron fundamentales para el desarrollo de este trabajo. Su compromiso con la excelencia académica y su disposición para compartir conocimiento dejaron una huella significativa en mi formación.

A la institución que me abrió sus puertas,
Gracias por brindar un entorno de aprendizaje estimulante, por fomentar el pensamiento crítico y por impulsar mi crecimiento profesional y personal. El respaldo y la motivación recibidos a lo largo de este proceso fueron clave para alcanzar este objetivo.

RESUMEN

El presente estudio de investigación aborda el desarrollo de un modelo de colorización automática de imágenes en escala de grises mediante un enfoque híbrido de aprendizaje profundo, que combina las fortalezas de tres arquitecturas: redes neuronales convolucionales (CNN), redes generativas adversariales (GAN) y *Vision Transformers* (ViT). El sistema propuesto tiene como objetivo reconstruir los colores perdidos en imágenes monocromáticas de manera realista, eficiente y coherente, apoyándose en la capacidad de aprendizaje de patrones espaciales locales, generación adversarial de detalles visuales y atención global a largo alcance.

Para su entrenamiento, se utilizó el conjunto de datos ImageNet, ampliamente reconocido por su variedad de clases y riqueza visual, lo cual permitió al modelo generalizar de forma adecuada a distintos contextos y estilos de imagen. La evaluación se realizó mediante métricas cuantitativas como el PSNR (*Peak Signal-to-Noise Ratio*) y el SSIM (*Structural Similarity Index*), lo que permitió validar tanto la fidelidad del color como la preservación de la estructura visual en las imágenes generadas.

Además, el modelo fue diseñado para adaptarse tanto a entornos con CPU (Unidad Central de Procesamiento) como GPU (Unidad de Procesamiento Gráfico), garantizando su versatilidad y aplicabilidad en diversas condiciones computacionales. Como conclusión, el sistema demostró ser efectivo en la tarea de colorización automática, superando modelos base en términos de calidad visual. Este trabajo representa un aporte significativo en el campo de la visión por computadora, con potencial para ser aplicado en la restauración de archivos históricos, arte digital y medios audiovisuales.

ABSTRACT

This research study presents the development of an automatic image colorization model for grayscale images using a hybrid deep learning approach that combines the strengths of three architectures: Convolutional Neural Networks (CNN), Generative Adversarial Networks (GAN), and Vision Transformers (ViT). The proposed system aims to realistically, efficiently, and coherently reconstruct the lost colors in monochromatic images by leveraging spatial pattern learning, adversarial detail generation, and long-range global attention.

The model was trained on the ImageNet dataset, widely recognized for its variety of classes and visual richness. This allows the system to generalize effectively across different image contexts and styles. The model was evaluated using quantitative metrics such as PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index), allowing for the validation of both color fidelity and structural consistency in the generated images.

Moreover, the model was designed to run on CPU and GPU environments, ensuring versatility across different computational settings. In conclusion, the system proved effective for automatic colorization tasks, outperforming baseline models in visual quality. This work represents a significant contribution to computer vision, with potential applications in image restoration, digital art, and audiovisual processing.

Índice general

Page

LISTA DE FIGURAS

Page

LISTA DE TABLAS

Page

Capítulo 1

Introducción

La colorización de imágenes ha cobrado un interés creciente en el ámbito de la visión por computadora debido a su potencial aplicación en la restauración de material histórico, la producción audiovisual y otras áreas donde la información cromática resulta fundamental. La presente tesis explora el uso de redes neuronales profundas para abordar este problema, considerando los avances recientes en aprendizaje profundo y en el procesamiento de imágenes.

El objetivo principal es desarrollar un sistema automatizado capaz de asignar valores de color de manera precisa y realista, enfocándose en la reconstrucción de detalles esenciales y la coherencia cromática general de la imagen.

1.1. Antecedentes

La colorización de imágenes es un problema ampliamente estudiado en el campo de la visión por computadora, con aplicaciones en la restauración de fotografías históricas, preservación de archivos y generación de contenido visual (Zhang et al., 2016). Tradicionalmente, este

proceso se realizaba de manera manual por artistas o restauradores, quienes aplicaban colores basándose en su conocimiento y referencias históricas. Sin embargo, este método era altamente subjetivo, costoso y demandaba una gran cantidad de tiempo, además de depender completamente de la interpretación del especialista (Levin et al., 2004).

Con el desarrollo de la inteligencia artificial y el aprendizaje profundo, han surgido modelos automatizados capaces de predecir los colores de imágenes en escala de grises, lo que ha permitido una colorización más rápida y consistente (Zhang et al., 2016). Los enfoques basados en redes neuronales han demostrado ser eficaces, aunque enfrentan desafíos como la interpretación contextual del color y la coherencia de las tonalidades generadas (Cheng et al., 2015). La comparación entre métodos tradicionales y los basados en aprendizaje profundo ha mostrado que, si bien la intervención humana sigue siendo necesaria en algunos casos, los modelos automatizados ofrecen una alternativa eficiente y escalable con resultados cada vez más realistas (Iizuka et al., 2016). La evolución de los métodos de colorización ha pasado de técnicas semánticas apoyadas en imágenes de referencia (Chia et al., 2011) a métodos completamente automáticos basados en redes neuronales profundas.

Además, han surgido enfoques multimodales que incorporan texto como guía para la colorización (Kim et al., 2019), lo cual abre nuevas posibilidades creativas, especialmente en contextos de animación y diseño digital. Otros trabajos han buscado explotar entradas como bosquejos o trazos de color para permitir una colorización interactiva y controlada, como lo demuestra el modelo Scribbler propuesto por Sangkloy et al. (2017), asi como tambien se han explorado técnicas basadas en perfiles de gradiente para preservar bordes y contornos (Cho et al., 2012).

1.1.1. Enfoques en Colorización Automática: Estudios Previos

Dentro de la colorización automática, se han explorado diversas metodologías y técnicas. Las Redes Neuronales Convolucionales (CNNs) han sido uno de los primeros enfoques utilizados para predecir colores basados en la estructura local de las imágenes. Zhang et al. (2016) aplicaron CNNs para entrenar un modelo que colorea imágenes a partir de la información de luminancia, logrando resultados satisfactorios en términos de precisión del color. Sin embargo, este enfoque tiene limitaciones en la coherencia global de la imagen, ya que las CNNs solo analizan características locales y no capturan relaciones a largo alcance dentro de la imagen.

Trabajos también han demostrado el potencial del aprendizaje, por ejemplo donde el modelo toma una imagen de referencia para inferir colores compatibles (Xu et al., 2020). Este tipo de enfoques resulta útil cuando se desea mantener una coherencia estilística específica o cuando se restauran imágenes con un contexto histórico o artístico definido.

En años recientes, la colorización automática de imágenes en escala de grises ha evolucionado drásticamente gracias a los avances en redes neuronales profundas. Técnicas como las Redes Generativas Adversariales (GAN) han demostrado una capacidad impresionante para aprender mapas de color coherentes y realistas a partir de datos de entrenamiento (Vitoria et al., 2020). Además, se han desarrollado enfoques que incorporan guía del usuario para controlar los resultados, como se muestra en el trabajo de Zhang et al. (2017), lo cual resulta útil en escenarios creativos o restaurativos.

Para mejorar la calidad de la colorización, las Redes Generativas Adversariales (GANs) fueron introducidas en este campo por Goodfellow et al. (2014) y posteriormente aplicadas por Larsson et al. (2016) y Radford et al. (2016). Estas redes emplean un generador y un discriminador para producir imágenes con colores más realistas, generando detalles que las CNNs por sí solas no pueden replicar. Las investigaciones recientes han demostrado que la

incorporación de información semántica más fina, como las instancias de objetos dentro de una escena, mejora notablemente la precisión de la colorización. Su et al. (2020) propusieron un enfoque *instance-aware* que distingue entre objetos similares para aplicar colores más coherentes en contextos complejos. En esta misma línea, el trabajo de Vitoria et al. (2020) con ChromaGAN implementa una arquitectura adversarial que aprende a distribuir los colores en función de la semántica de clase, logrando una colorización más realista, especialmente en imágenes con múltiples elementos o escenas urbanas. Estos avances reflejan una clara tendencia hacia modelos que no solo colorean, sino que comprenden el contenido visual en un nivel más profundo. No obstante, las GANs también presentan desafíos, como la inestabilidad en el entrenamiento y la generación de artefactos visuales, lo que ha llevado a la exploración de modelos híbridos.

Más recientemente, los *Vision Transformers* (ViTs) han sido explorados como una alternativa prometedora en el procesamiento de imágenes. Dosovitskiy et al. (2021) introdujeron esta arquitectura, basada en el mecanismo de atención de los *Transformers*, con el objetivo de capturar relaciones globales en las imágenes. Su implementación en colorización sigue siendo un área en desarrollo, pero su capacidad para mantener la coherencia del color en regiones extensas lo convierte en un candidato interesante para mejorar los resultados obtenidos con CNNs y GANs.

1.1.2. Integración de Algoritmos para Mejorar la Colorización de Imágenes

Debido a las limitaciones de los enfoques tradicionales, han surgido propuestas que combinan diferentes modelos para aprovechar las fortalezas de cada uno. Kumar et al. (2024) propusieron *ParaColorizer: Realistic Image Colorization using Parallel Generative Networks* propone un marco basado en redes generativas adversariales (GANs) paralelas, donde una

red se encarga de la colorización del primer plano utilizando características a nivel de objeto, mientras que otra red se enfoca en el fondo basado en características de la imágen completa. La fusión de estos resultados mediante una red basada en DenseFuse logra una colorización más realista y coherente. Otro enfoque híbrido es el presentado por Baldassarre et al. (2017) en *Deep Koalarization: Image Colorization using CNNs and Inception-ResNet-v2*, donde los autores combinaron una CNN entrenada desde cero con características de alto nivel extraídas del modelo Inception-ResNet-v2. Esta arquitectura de *encoder-decoder* convolucional permite procesar imágenes de diferentes tamaños y proporciones, mejorando la precisión y calidad de la colorización. En la línea de modelos combinados, Ragab et al. (2023) propusieron *Incorporating Ensemble and Transfer Learning For An End-To-End Auto-Colorized Image Detection Model* implementa técnicas de *ensemble learning* y aprendizaje por transferencia. Utiliza múltiples arquitecturas preentrenadas, como VGG16, ResNet50, MobileNet v2 y EfficientNet, para mejorar la clasificación y detección de imágenes colorizadas automáticamente. Aunque este modelo se centra en la identificación de imágenes generadas por computadora, su metodología destaca el uso de combinaciones de redes neuronales para mejorar el rendimiento del modelo. Estos estudios evidencian que la combinación de diferentes algoritmos puede mejorar la calidad de la colorización de imágenes, abordando las deficiencias de enfoques individuales y permitiendo generar colores más realistas y con mejor coherencia global.

El presente trabajo surge de la necesidad de mejorar la calidad de la colorización automática mediante la integración de múltiples técnicas de aprendizaje profundo, combinando CNNs, GANs y ViTs para aprovechar las ventajas de cada uno. A diferencia de estudios previos que han explorado estos modelos de manera independiente, este proyecto propone un modelo híbrido que busca mejorar la coherencia global y el realismo del color generado.

En cuanto a la evaluación de los resultados, métricas como PSNR (*Peak Signal-to-Noise Ratio*) y SSIM (*Structural Similarity Index*) siguen siendo ampliamente utilizadas para cuan-

tificar la similitud entre la imagen original y la imagen colorizada (Wang et al., 2003), así como comparaciones cualitativas con trabajos previos (Wang et al., 2004; Zhang et al., 2016). Con este enfoque, se espera obtener una solución innovadora que reduzca las inconsistencias cromáticas y ofrezca una alternativa efectiva en el proceso de colorización de imágenes (Zhang et al., 2016).

1.2. Planteamiento del problema

En la actualidad, existe una amplia variedad de situaciones donde la información de color en una imagen puede ser limitada, deficiente o inadecuada (Cheng et al., 2015). Los métodos manuales para la colorización suelen ser lentos, costosos y altamente dependientes de la pericia del artista o restaurador (Levin et al., 2004). Asimismo, la ambigüedad inherente al color (puesto que varios tonos podrían ser válidos para una misma región de la imagen) complica la tarea de asignar colores adecuados sin un criterio objetivo.

Desde la perspectiva de la inteligencia artificial, las redes neuronales profundas ofrecen una vía para automatizar este proceso, aprendiendo a partir de grandes colecciones de imágenes a color. Sin embargo, los modelos deben tratar con la variabilidad de las texturas, la iluminación y los contextos semánticos presentes en cada imagen, lo cual representa un desafío significativo para el entrenamiento y la generalización (Zhang et al., 2016).

1.3. Justificación

La colorización automática de imágenes se justifica por su amplio potencial en diversos campos. En el ámbito de la conservación y restauración de archivos históricos, esta técnica permite enriquecer registros fotográficos que carecen de información cromática, aumentando

su valor documental y atractivo visual, así como facilitando su preservación digital. En las industrias creativas, como la cinematografía y la producción audiovisual, la posibilidad de añadir o modificar colores de manera automática abre nuevas oportunidades en la postproducción y la generación de contenido innovador. Además, el acceso cada vez más extendido a herramientas de aprendizaje profundo y recursos computacionales impulsa el desarrollo de soluciones más precisas, accesibles y escalables, que reducen significativamente los costos y tiempos de procesamiento. Finalmente, este tipo de aplicaciones fomenta la investigación científica al integrar problemáticas complejas relacionadas con el reconocimiento de patrones, el aprendizaje semántico y la generación de contenido, lo cual contribuye al avance de las técnicas de visión por computadora (Zhang et al., 2016).

1.4. Preguntas de investigación

1. ¿Qué tan efectiva es la integración de CNN, GAN y Vision Transformers (ViT) en un modelo híbrido para la colorización automática de imágenes?

2. ¿Cómo influye la selección y el tamaño del conjunto de datos en la calidad de la colorización?

3. ¿Qué métodos de evaluación (métricas objetivas y subjetivas) pueden emplearse para medir el realismo y la coherencia de los colores generados?

4. ¿Son suficientes los indicadores actuales para comparar el modelo propuesto con los modelos tradicionales de colorización de imágenes?

1.5. Hipótesis

- Si se entrena un modelo híbrido basado en redes neuronales profundas con un conjunto de datos amplio y diverso, entonces será capaz de generar imágenes colorizadas con precisión cromática y coherencia contextual similares a la percepción humana, lo cual se comprobará mediante métricas cuantitativas como el PSNR (*Peak Signal-to-Noise Ratio*) y el SSIM (*Structural Similarity Index*). Estas métricas permitirán evaluar la fidelidad del color generado y la similitud estructural entre la imagen original y la imagen colorizada desde un enfoque objetivo.

1.6. Objetivo general

Diseñar y evaluar un modelo híbrido que combine redes neuronales convolucionales (CNN), modelos generativos adversariales (GAN) y *Vision Transformers* (ViTs), que mejorará la colorización automática de imágenes, obteniendo resultados más precisos en términos de coherencia contextual y percepción cromática en comparación con métodos tradicionales.

1.7. Objetivos específicos

- Revisar el estado del arte en métodos de colorización de imágenes con énfasis en redes neuronales y técnicas de aprendizaje profundo.
- Seleccionar y procesar un conjunto de datos que permita el entrenamiento, la validación y la prueba del modelo de manera efectiva, analizando su impacto en la calidad de los resultados generados.
- Desarrollar y entrenar un prototipo de red neuronal híbrida, explorando arquitecturas basadas en redes neuronales convolucionales (CNN), modelos generativos adversariales

(GAN) y *Vision Transformers* (ViT), con el fin de evaluar su efectividad en la tarea de colorización automática.

- Implementar y comparar diferentes estrategias de entrenamiento, incluyendo el uso de espacios de color, funciones de pérdida y esquemas de regularización.

- Analizar el desempeño del modelo mediante métricas objetivas (p. ej., PSNR (*Peak Signal-to-Noise Ratio*), SSIM (*Structural Similarity Index*)) y la evaluación subjetiva de la calidad visual por parte de expertos o usuarios finales, para medir el realismo y coherencia cromática.

- Proponer mejoras e identificar aplicaciones futuras, considerando la escalabilidad y la adaptabilidad del sistema, así como su capacidad para superar los métodos tradicionales.

Capítulo 2

Marco Teórico

La colorización de imágenes ha adquirido gran relevancia en el campo de la visión por computadora debido a su potencial aplicación en la restauración de material visual, la producción audiovisual y otros ámbitos donde la información cromática resulta esencial (Cheng et al., 2015; Zhang et al., 2016). Este proyecto se enfoca en los sistemas de colorización automática de imágenes mediante técnicas de aprendizaje profundo, específicamente aquellos que integran redes neuronales convolucionales (CNN), redes generativas adversariales (GAN) y transformadores de visión (ViT, por sus siglas en inglés). A lo largo de este proyecto se abordan los principales avances técnicos y académicos que han impulsado el desarrollo de modelos híbridos más precisos, capaces de generar imágenes colorizadas de alta calidad a partir de entradas en escala de grises.

Se revisan conceptos clave como la convolución, las funciones de activación, las capas de normalización y agrupamiento, así como los mecanismos de atención y la arquitectura *encoder-decoder*. Además, se analiza la evolución de los enfoques clásicos hacia modelos más complejos basados en aprendizaje profundo, destacando las ventajas de la combinación de múltiples arquitecturas. También se incluyen los conjuntos de datos más relevantes utilizados para

entrenamiento y evaluación, así como las métricas empleadas para medir el rendimiento de los sistemas, tales como PSNR (*Peak Signal-to-Noise Ratio*) y SSIM (*Structural Similarity Index*).

El objetivo principal es desarrollar un modelo automatizado que asigne, recupere o modifique valores cromáticos de manera coherente y realista, garantizando una representación visual precisa de la información de color en cada imagen.

2.1. Revisión de la literatura

La colorización de imágenes ha sido objeto de investigación desde mediados del siglo XX, inicialmente abordada mediante técnicas manuales o basadas en reglas heurísticas. Sin embargo, el desarrollo del aprendizaje profundo ha transformado radicalmente este campo. A partir de 2016, trabajos como los de Zhang et al. (2016). introdujeron modelos basados en CNNs que lograban colorizar imágenes en baja resolución con resultados visualmente aceptables. Estos modelos se entrenaban utilizando grandes conjuntos de imágenes RGB convertidas al espacio *CIELAB*, prediciendo los canales de color (a, b) a partir del canal de luminancia (L). Aunque innovadores, estos enfoques enfrentaban limitaciones relacionadas con la saturación del color y la falta de contexto global, lo que afectaba la coherencia cromática en escenas complejas Zhang et al. (2016).

Posteriormente, las redes GAN, introducidas por Goodfellow et al. (2014), fueron adoptadas en el contexto de colorización para generar resultados más realistas. Modelos como Pix2Pix y sus derivados demostraron que un discriminador podía guiar al generador para producir colores más coherentes y naturales. No obstante, estos modelos presentaban problemas de estabilidad durante el entrenamiento y requerían ajustes precisos en la función de pérdida.

Con la introducción de los *Vision Transformers* (ViT), el campo experimentó una nueva ola

de avances, al reemplazar las convoluciones tradicionales por mecanismos de atención que capturan relaciones globales en la imagen desde las primeras capas del modelo (Dosovitskiy et al., 2021). Al aprovechar mecanismos de atención auto-regresiva, los ViT permitieron capturar dependencias espaciales globales en las imágenes, mejorando la contextualización y la coherencia semántica del color. Sin embargo, debido a su alta demanda computacional, los ViT puros no eran ideales para sistemas en tiempo real o para ejecución en *hardware* limitado. Como resultado, investigaciones recientes se han centrado en modelos híbridos que combinan CNNs para el procesamiento local eficiente, GANs para la generación realista, y ViT para la contextualización global, logrando un balance entre rendimiento, realismo y eficiencia (Lugmayr et al., 2020; Ali et al., 2023).

2.1.1. Sistemas de colorización de imágenes

Los primeros enfoques de colorización automática con *deep learning* empleaban CNNs como arquitectura base. Un ejemplo temprano es el modelo de Zhang et al. (2016), que usaba una red *encoder-decoder* para predecir los canales a y b del espacio CIELAB. Aunque efectivo en la reconstrucción de colores básicos, este modelo tendía a generar imágenes con colores apagados y problemas de ambigüedad semántica. Para abordar esta limitación, Iizuka et al. (2016) propusieron una arquitectura dual que combinaba ramas locales y globales, logrando mejores resultados contextuales.

La introducción de GANs marcó un hito, ya que permitieron obtener imágenes con colores más vivos y realistas. En el modelo Pix2Pix, por ejemplo, el generador es entrenado adversarialmente contra un discriminador que aprende a distinguir imágenes reales de las generadas. Esto incentivó la generación de resultados más fotorealistas, aunque también introdujo desafíos como el colapso del modo y la necesidad de ajustes finos en la arquitectura y función de pérdida (Isola et al., 2017).

Recientemente, los *Vision Transformers* han sido explorados por su capacidad para captar información a nivel global. En 2021, Esser et al. (2021) presentaron el modelo TransColor, un ViT adaptado a tareas de colorización, que logró resultados superiores en tareas de recuperación semántica del color. No obstante, su alto costo computacional motivó la búsqueda de enfoques híbridos.

Otros trabajos recientes han explorado la inclusión de módulos de atención jerárquica, redes de realce perceptual, y funciones de pérdida compuestas por PSNR, SSIM y pérdida perceptual basada en redes como VGG (*Visual Geometry Group*). Estos enfoques buscan mejorar tanto la fidelidad objetiva (medida por PSNR) como la calidad subjetiva (valorada por usuarios).

Además, la elección del conjunto de datos tiene un papel crucial en el entrenamiento de estos modelos. Conjuntos como CIFAR-10, ImageNet, Places365 y Tiny ImageNet han sido ampliamente utilizados, siendo este último particularmente valioso para tareas de validación ligera debido a su tamaño reducido y diversidad de clases. Sin embargo, también se ha propuesto el uso de datasets con información semántica y médica como el conjunto histológico publicado por Kather et al. (2019), para probar la robustez de los modelos en imágenes más especializadas. Para asegurar un entrenamiento robusto, se emplean técnicas de normalización, aumento de datos, y balanceo de clases, que ayudan a mejorar la generalización del modelo.

Finalmente, las métricas de evaluación más comunes en este dominio incluyen PSNR, SSIM, LPIPS (*Learned Perceptual Image Patch Similarity*), y FID (*Fréchet Inception Distance*). Estas métricas permiten cuantificar tanto la similitud estructural como la percepción subjetiva de los colores generados, brindando un marco objetivo para comparar el rendimiento entre diferentes modelos (Wang et al., 2004; Zhang et al., 2018).

2.2. Estado del arte: investigaciones previas en colorización de imágenes

El aprendizaje profundo (*deep learning*, DL) es una subdisciplina del aprendizaje automático que se inspira en la estructura y el funcionamiento del cerebro humano, particularmente en el comportamiento de las neuronas biológicas (Goodfellow et al., 2014).

Inspirado en las neuronas biológicas, el modelo de una neurona artificial simula la recepción de señales de entrada, su procesamiento mediante una función de activación y la generación de una salida. Formalmente, para un vector de entradas $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ y un vector de pesos $\mathbf{w} = [w_1, w_2, \ldots, w_n]$, la salida $\hat{y}$ se calcula como:

$$\hat{y} = g\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{2.1}$$

donde b representa el sesgo (*bias*) y g es la función de activación, que introduce no linealidad al modelo. Algunas funciones de activación comunes incluyen:

- **ReLU (*Rectified Linear Unit*)**: Es una función no lineal que produce una salida de cero para valores negativos de entrada y una salida lineal para valores positivos. Reduciendo significativamente el problema del desvanecimiento del gradiente en comparación con funciones como la sigmoide o tanh, dando mejora a la eficiencia computacional durante el entrenamiento, lo que contribuye a una convergencia más rápida en redes profundas (Nair and Hinton, 2010). Aunque tambien se debe tener cuidado ya que puede causar el fenómeno conocido como "neuronas muertas", cuando algunos nodos dejan de activarse completamente y ya no contribuyen al aprendizaje (Maas

et al., 2013). La función ReLU se define matemáticamente en la Ecuación (2.2):

$$g(z) = \text{máx}(0, z) \tag{2.2}$$

donde z representa la suma ponderada de las entradas de la neurona más el término de sesgo

- **Sigmoide**: Ayuda a mapear cualquier valor real a un rango entre 0 y 1. Su salida puede interpretarse como una probabilidad, siendo útil en tareas de clasificación binaria, especialmente en la capa de salida, que proporciona una transición suave y continua entre valores. En redes profundas, puede causar problemas de saturación y desvanecimiento del gradiente, lo que ralentiza o impide el aprendizaje (Bishop, 2006). No obstante, su uso en redes profundas puede presentar inconvenientes. Entre ellos destacan la saturación de los gradientes (cuando los valores de entrada son muy grandes o muy pequeños, la derivada tiende a cero) y el desvanecimiento del gradiente, lo que puede ralentizar o incluso impedir el aprendizaje efectivo del modelo (Glorot and Bengio, 2010). La función sigmoide se define matemáticamente en la Ecuación (2.3):

$$g(z) = \frac{1}{1 + e^{-z}} \tag{2.3}$$

- **Tanh**: Es una versión escalada y centrada de la sigmoide, mapeando la entrada al rango [-1, 1], a diferencia de la sigmoide que lo hace en [0, 1]. Mejora el aprendizaje al estar centrada en cero, por lo que tiende a producir una mejor convergencia que la sigmoide (LeCun et al., 1998), útil en capas ocultas cuando se requiere una activación que permita tanto valores positivos como negativos. La función tanh se define matemáticamente en la Ecuación (2.4):

$$g(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}} \tag{2.4}$$

Estas funciones permiten a la red neuronal modelar relaciones no lineales complejas entre las variables de entrada y salida (Goodfellow et al., 2014).

Estas redes pueden organizarse en una sola capa (*Single-Layer Perceptron*, SLP) o en múltiples capas (*Multi-Layer Perceptron*, MLP), permitiendo la modelación de patrones complejos y no lineales (LeCun et al., 2015).

Los avances recientes en aprendizaje profundo han posibilitado el desarrollo de modelos sofisticados aplicables a tareas complejas como la colorización automática de imágenes. En particular, las redes neuronales convolucionales (CNN) permiten capturar patrones espaciales locales a través de filtros entrenables, lo que las convierte en una herramienta esencial para el procesamiento de imágenes (Krizhevsky et al., 2017). Por otro lado, las redes generativas adversariales (*Generative Adversarial Networks*, GAN) permiten generar imágenes realistas mediante un proceso de competencia entre un generador y un discriminador (Goodfellow et al., 2014).

Los avances recientes en aprendizaje profundo han posibilitado el desarrollo de modelos sofisticados aplicables a tareas complejas como la colorización automática de imágenes. En particular, las redes neuronales convolucionales (CNN) permiten capturar patrones espaciales locales a través de filtros entrenables, lo que las convierte en una herramienta esencial para el procesamiento de imágenes (Krizhevsky et al., 2017). Por otro lado, las redes generativas adversariales (GAN) permiten generar imágenes realistas mediante un proceso de competencia entre un generador y un discriminador (Goodfellow et al., 2014).

Más recientemente, los modelos tipo *transformer*, inicialmente diseñados para procesamiento del lenguaje natural, han sido adaptados para tareas de visión computacional. Los *Vision Transformers* (ViT) permiten modelar relaciones globales entre píxeles de una imagen a través de mecanismos de atención auto-regulada, aportando contexto global al proceso de generación o transformación de imágenes (Dosovitskiy et al., 2021).

La combinación de estas tres arquitecturas (CNN + GAN + ViT) permite aprovechar sus respectivas fortalezas: la capacidad local de las CNN, la generación realista de las GAN, y el modelado contextual global de los transformers. En el contexto de este trabajo, se propone una arquitectura híbrida para realizar la colorización automática de imágenes en escala de grises, optimizando la fidelidad visual y la coherencia contextual del color.

La colorización automática de imágenes ha sido objeto de estudio en diversas investigaciones. Entre las principales metodologías empleadas, destacan:

- **Redes Neuronales Convolucionales (CNNs)**: (Zhang et al., 2016) desarrollaron un modelo basado en CNNs para predecir los valores cromáticos a partir de la estructura de la imagen. Si bien esta técnica permite obtener resultados aceptables, en ocasiones presenta problemas de inconsistencia cromática. Iizuka et al. (2016) propusieron un modelo híbrido basado en CNNs que incorpora información contextual global, mejorando la coherencia de la colorización en escenas más complejas. Asi como He et al. (2016) impulsa el aprendizaje de patrones locales y la reconstrucción de la información cromática de forma pixelada.

  En la práctica, las CNN han demostrado ser muy útiles en tareas como:

  - Reconocimiento de imágenes y videos (ej: clasificación de fotos, detección de rostros).
  - Procesamiento de lenguaje natural (NLP) en algunos casos.
  - Diagnóstico médico (análisis de radiografías).
  - Autos autónomos (detección de peatones, señales de tránsito).

  La Figura 2.1 ilustra la arquitectura general de una red neuronal convolucional, destacando sus componentes principales y el flujo de procesamiento de datos.

- **Redes Adversariales Generativas (GANs)**: Goodfellow et al. (2014) introdujeron

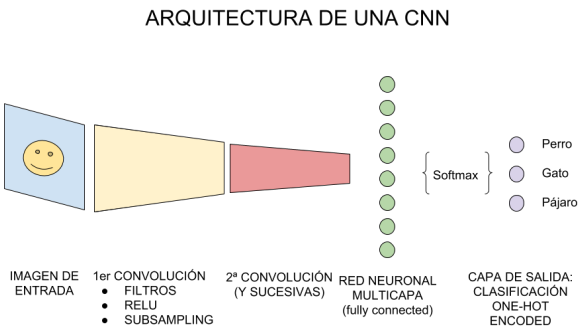


Figura 2.1: Diagrama de la arquitectura de la red neuronal convolucional (CNN) utilizada en este trabajo, ilustrando las principales capas y el flujo de datos desde la entrada hasta la salida.

las GANs como un enfoque para la generación de imágenes, permitiendo que modelos de aprendizaje profundo generen representaciones más realistas. Larsson et al. (2016) aplicaron esta técnica en la colorización de imágenes, logrando una mayor naturalidad en la generación de colores. Además, Zhang et al. (2016) desarrollaron Pix2Pix, un modelo basado en GANs para la conversión de imágenes en diferentes dominios, incluyendo la colorización, obteniendo mejoras significativas en la calidad del color generado.

Las GAN se han aplicado en muchos campos como:

- Generación de imágenes de personas que no existen
- Restauración de fotos antiguas o en baja resolución
- Diseño de ropa o muebles con ejemplos visuales
- Creación de arte digital

La Figura 2.2 muestra la estructura básica de una GAN, incluyendo el flujo de datos entre el generador y el discriminador.

- ***Transformers* en Visión Artificial**: Recientemente, Dosovitskiy et al. (2021) presentaron *Vision Transformers* (ViTs), una arquitectura que mejora la captación de

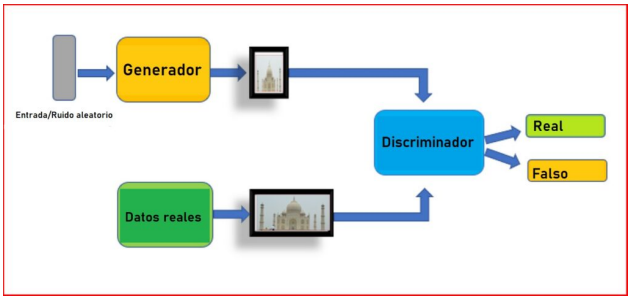


Figura 2.2: Arquitectura general de una red generativa adversarial (GAN), que consta de un generador y un discriminador en competencia.

patrones globales en imágenes, lo que podría beneficiar la colorización al mantener la coherencia de la información cromática en grandes regiones de la imagen. Su capacidad de capturar relaciones globales en la imagen puede complementar la generación de color realizada por GANs, evitando errores comunes en CNNs tradicionales. Saharia et al. (2022) propusieron Palette, un modelo basado en *Transformers* diseñado específicamente para tareas de manipulación de imágenes, incluyendo la colorización, logrando resultados con una mejor distribución del color y menos artefactos visuales.

Esta capacidad de "ver el todo" hace que los ViT sean muy buenos en:

- Clasificación de imágenes complejas
- Detección de anomalías en imágenes industriales
- Análisis de escenas completas (por ejemplo, una imagen de tráfico)

La Figura 2.3 ilustra el funcionamiento general de esta arquitectura aplicada a visión por computadora.

Estos estudios han permitido mejorar la precisión de la colorización, aunque aún existen áreas de oportunidad en cuanto a la preservación de texturas, la generalización a distintos tipos de imágenes y la interpretación contextual de los colores. La combinación de estos enfoques en un modelo híbrido, como el propuesto en este proyecto, busca aprovechar las fortalezas de cada uno para mejorar la calidad de la colorización automática.

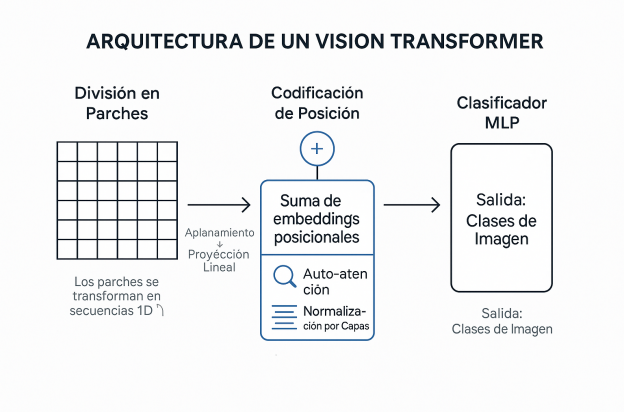


Figura 2.3: Arquitectura simplificada de un *Vision Transformer* (ViT), que convierte una imagen en parches, los proyecta como *embeddings* y los procesa mediante bloques tipo *Transformer*.

2.3. Redes Neuronales Convolucionales (CNNs)

Las CNNs son arquitecturas profundas que extraen características espaciales jerárquicas mediante capas convolucionales. Son altamente efectivas en tareas visuales como clasificación y segmentación.

2.3.1. Componentes principales

- **Capas convolucionales** ($Conv2D$):

  Las capas convolucionales son el núcleo de las CNNs. Su función es aplicar un conjunto de filtros (también llamados *kernels*) sobre la imagen de entrada para extraer características locales como bordes, texturas, patrones o formas. Cada filtro se desliza sobre la entrada realizando una operación matemática conocida como convolución, generando un mapa de activación (*feature map*) que resalta las regiones donde el filtro detecta una característica específica.

  Matemáticamente, la convolución entre una entrada I y un filtro K se representa como:

$$S_{i,j} = (I * K)_{i,j} = \sum_m \sum_n I_{i+m,j+n} \cdot K_{m,n} \tag{2.5}$$

A medida que se avanza a través de las capas convolucionales, la red aprende representaciones cada vez más abstractas y de mayor nivel. Este proceso jerárquico es fundamental para entender contextos complejos en imágenes (Krizhevsky et al., 2017).

- **Funciones de activación (ReLU)**:

  Tras cada operación convolucional, se aplica una función de activación para introducir no linealidad al modelo, permitiéndole aprender relaciones complejas. La función más comúnmente usada en CNNs es la ReLU (*Rectified Linear Unit*), definida en la Ecuación (2.2).

  ReLU mejora significativamente la eficiencia del entrenamiento al evitar la saturación de gradientes y acelerar la convergencia, comparada con funciones como la sigmoide o *tanh* (Nair and Hinton, 2010). Además, al establecer a cero las salidas negativas, promueve la activación dispersa, lo que mejora la generalización.

- **Pooling (*Max/Avg Pooling*)**:

  El *pooling* es una operación de reducción de dimensionalidad que se aplica para disminuir el tamaño espacial de los mapas de activación, mantener las características más relevantes y reducir el costo computacional. Los dos tipos más comunes son:

  - ***Max Pooling*:** selecciona el valor máximo dentro de una ventana (por ejemplo, 2x2) y lo propaga al siguiente nivel.
  - ***Average Pooling*:** calcula el promedio de los valores en la ventana.

  El *pooling* no sólo contribuye a la eficiencia, sino que también introduce invarianza a pequeñas transformaciones, como traslaciones o distorsiones locales en la imagen (Scherer et al., 2010). Esta propiedad es valiosa en aplicaciones de visión por computadora.

- **Capas totalmente conectadas (*Fully Connected*)**:

  Las capas totalmente conectadas, también conocidas como *dense layers*, son aquellas donde cada neurona se conecta con todas las neuronas de la capa anterior. Estas capas suelen ubicarse al final de la red y son responsables de la interpretación final de las características extraídas por las capas convolucionales y de *pooling*.

  En el caso de tareas de clasificación, estas capas producen las probabilidades de pertenencia a cada clase mediante funciones como *softmax*. En tareas de regresión (como la predicción de canales de color en colorización), la salida puede ser un vector de valores continuos (Zhang et al., 2016).

  Aunque son efectivas para la integración global de la información, tienden a tener muchos parámetros y, por tanto, son propensas al sobreajuste. Es común aplicar técnicas como regularización o *dropout* para mitigar este problema.

2.3.2. Formulación matemática básica

Las Redes Neuronales Convolucionales (CNN, por sus siglas en inglés) fueron desarrolladas para procesar datos con estructura de cuadrícula, como imágenes. La operación principal en una CNN es la convolución, definida en la Ecuación (2.5).

Las CNN utilizan múltiples filtros para aprender características como bordes, texturas y formas, que luego se combinan en capas profundas para reconocer estructuras más complejas (LeCun et al., 1998).

2.3.3. Arquitectura típica

Las redes neuronales convolucionales (CNN) se estructuran a partir de bloques repetitivos que combinan capas convolucionales, funciones de activación y operaciones de reducción

de dimensionalidad (*pooling*). En el contexto de la colorización automática de imágenes, se suelen emplear arquitecturas especializadas como VGGNet (Simonyan and Zisserman, 2015), ResNet (He et al., 2016) y U-Net (Ronneberger et al., 2015), las cuales se han consolidado por su efectividad en la extracción y reconstrucción de características espaciales relevantes.

VGGNet se caracteriza por su estructura profunda y homogénea, basada exclusivamente en filtros pequeños de tamaño 3×3, lo que permite una captura progresiva de patrones visuales desde los más simples a los más complejos (Simonyan and Zisserman, 2015).

ResNet introduce conexiones residuales que permiten el flujo directo de la información a través de la red, lo que facilita el entrenamiento de arquitecturas muy profundas sin sufrir problemas de desvanecimiento del gradiente (He et al., 2016).

Por su parte, U-Net es una arquitectura *encoder-decoder* que incluye conexiones de tipo *skip*, las cuales combinan características de baja y alta resolución, permitiendo una reconstrucción precisa del contenido espacial, lo que la hace particularmente útil en tareas de segmentación y colorización (Ronneberger et al., 2015).

En la colorización automática, estas arquitecturas suelen utilizarse en configuración *encoder-decoder*, donde el *encoder* reduce progresivamente la resolución espacial para capturar representaciones abstractas, y el *decoder* la reconstruye para generar los canales de color.

En la Figura 2.4 esquematiza una arquitectura típica utilizada en la colorización de imágenes. Se puede observar cómo la entrada en escala de grises es procesada a través de capas convolucionales y de reducción dimensional hasta alcanzar un espacio latente compacto. A partir de este espacio, el *decoder* reconstruye los canales cromáticos de la imagen, permitiendo una salida en color.

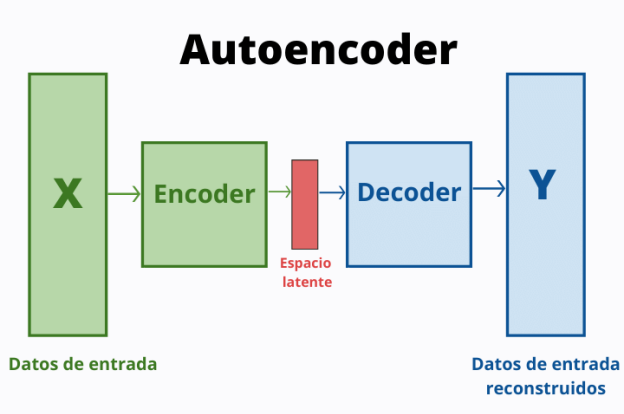


Figura 2.4: Ejemplo de arquitectura CNN tipo *encoder-decoder* para colorización automática. El *encoder* extrae características de la imagen en escala de grises, mientras que el *decoder* reconstruye los canales de color.

2.3.4. Hiperparámetros comunes

Los hiperparámetros controlan el comportamiento del modelo durante el entrenamiento y determinan en gran medida su rendimiento final. En redes CNN para tareas de colorización, algunos de los más comunes son los siguientes:

- **Tamaño del filtro:** 3×3

  Este tamaño se ha establecido como un estándar debido a su capacidad de capturar patrones espaciales locales sin aumentar excesivamente la complejidad computacional. Filtros pequeños permiten redes más profundas, facilitando una mayor expresividad del modelo sin comprometer la eficiencia (Simonyan and Zisserman, 2015).

- **Número de filtros por capa: 64, 128, 256**

  El número de filtros determina cuántas características diferentes puede aprender una capa. En etapas iniciales se suelen usar 64 filtros para extraer patrones básicos, mientras que en capas más profundas se emplean 128 o 256 para detectar estructuras más complejas como formas, texturas y bordes.

- **Función de activación: ReLU**

La función *Rectified Linear Unit* (ReLU), definida en la Ecuación (2.2), es ampliamente utilizada por su simplicidad y eficiencia computacional. ReLU introduce no linealidad al modelo, permitiendo que aprenda representaciones complejas, además de mitigar el problema del desvanecimiento del gradiente y acelerar el entrenamiento (Nair and Hinton, 2010).

- **Optimizador: Adam (tasa de aprendizaje 10^{-3})**

  Adam (*Adaptive Moment Estimation*) combina las ventajas de AdaGrad y RMSprop, realizando actualizaciones de pesos eficientes y adaptativas. Una tasa de aprendizaje inicial de 10^{-3} suele ser apropiada para tareas de colorización, permitiendo una convergencia estable sin oscilaciones bruscas (Kingma and Ba, 2014).

Estos hiperparámetros pueden ajustarse mediante validación cruzada o técnicas automáticas como búsqueda aleatoria o algoritmos bayesianos, dependiendo del problema específico.

2.3.5. Aplicación en colorización

En el ámbito de la colorización automática, las redes neuronales convolucionales se utilizan para aprender una función de mapeo entre la luminancia (canal **L** en el espacio de color Lab) y los componentes cromáticos (**a** y **b**). En otras palabras, dado un mapa de luminancia, el modelo CNN predice los valores correspondientes de color, completando así la información faltante.

Este enfoque es posible porque las CNN son capaces de aprender patrones espaciales complejos y relacionarlos con características de color observadas en el conjunto de entrenamiento. Por ejemplo, pueden aprender que los cielos son comúnmente azules o que las hojas suelen tener tonos verdes, asignando así colores contextualmente apropiados a regiones semánticas similares.

El uso del espacio Lab resulta estratégico, ya que separa la información de luminancia (brillo) de la información cromática, facilitando el aprendizaje de los colores sin que se vean afectados por cambios en la iluminación. Además, el canal L proporciona una estructura espacial detallada que guía al modelo en la reconstrucción de los colores.

Sin embargo, debido a la ambigüedad inherente del problema (una imagen en escala de grises puede corresponder a múltiples combinaciones de color válidas), la red debe apoyarse en el contexto global de la imagen y en patrones previamente aprendidos para realizar inferencias plausibles, lo cual resalta la importancia de arquitecturas profundas y *datasets* diversos.

2.4. Redes Generativas Adversariales (GANs)

Las GANs consisten en un generador G y un discriminador D en competencia. G genera muestras sintéticas; D distingue entre reales y falsas.

Las Redes Generativas Adversariales (GAN, por sus siglas en inglés) constan de dos redes neuronales enfrentadas:

- Un generador $G(z)$, que transforma un vector de ruido $z \sim p_z(z)$ en una imagen sintética $G(z)$.
- Un discriminador $D(x)$, que intenta distinguir entre imágenes reales $x \sim p_{\text{data}}(x)$ e imágenes generadas por el generador.

2.4.1. Función de pérdida adversarial

El entrenamiento de una GAN se formula como un juego de suma cero, definido por la siguiente función objetivo (Goodfellow et al., 2014):

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}\left[\log D(x)\right] + \mathbb{E}_{z\sim p_z(z)}\left[\log(1 - D(G(z)))\right] \tag{2.6}$$

Este enfoque ha demostrado ser capaz de generar imágenes altamente realistas. En tareas de colorización automática, se utiliza para entrenar al generador a producir imágenes colorizadas que sean difíciles de distinguir de imágenes reales (Isola et al., 2017).

2.4.2. Arquitectura

Las Redes Generativas Antagónicas (GANs) se componen de dos redes principales que compiten entre sí: el generador y el discriminador. Esta estructura de entrenamiento adversarial permite que el generador aprenda a producir imágenes cada vez más realistas, mientras que el discriminador aprende a diferenciar entre imágenes reales y generadas (Goodfellow et al., 2014).

- **Generador**: CNN estilo U-Net

  El generador en tareas de colorización generalmente adopta una arquitectura U-Net, una red convolucional de tipo *encoder-decoder* con conexiones de salto entre capas simétricas del codificador y del decodificador (Ronneberger et al., 2015). Estas conexiones permiten transferir detalles espaciales finos del codificador al decodificador, mejorando así la calidad de la reconstrucción de los colores. La U-Net es especialmente eficaz en tareas de segmentación y colorización porque retiene tanto información de contexto como detalles locales.

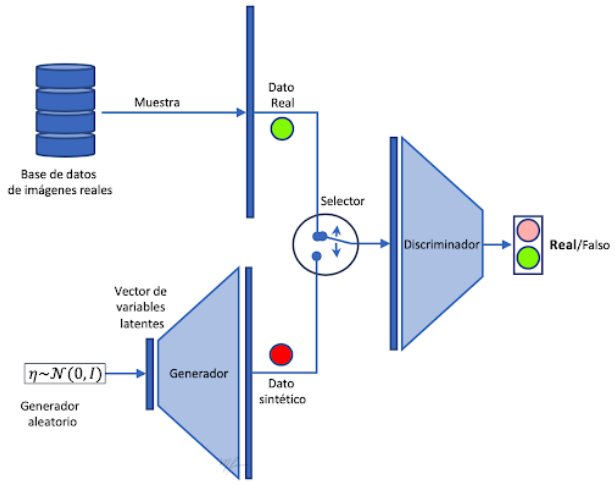


Figura 2.5: Arquitectura general de una GAN para colorización de imágenes. El generador recibe una imagen en escala de grises y produce una imagen colorizada. El discriminador compara esta salida con imágenes reales para distinguir cuál es falsa.

- **Discriminador**: CNN con arquitectura PatchGAN

  El discriminador empleado suele ser un PatchGAN, una red convolucional que no clasifica la imagen completa como real o falsa, sino que lo hace en bloques o parches, típicamente de 70×70 píxeles (Isola et al., 2017). Esta técnica permite centrarse en texturas locales, lo cual es crucial para asegurar la coherencia del color en pequeñas regiones de la imagen. El objetivo del discriminador es aumentar su precisión para distinguir entre imágenes realistas y generadas, forzando al generador a producir resultados más naturales.

La Figura 2.5 ilustra el flujo típico en una GAN para colorización. A la izquierda se muestra el generador tipo U-Net que transforma una imagen en escala de grises (canal L) en una imagen colorizada (predicción de los canales a y b). A la derecha, el discriminador PatchGAN evalúa localmente si los parches de la imagen generada se parecen a los de las imágenes reales, retroalimentando al generador para mejorar su desempeño.

2.4.3. Hiperparámetros comunes

Los hiperparámetros controlan el comportamiento y la eficacia del proceso de entrenamiento de una GAN. Los siguientes son algunos de los más utilizados en colorización de imágenes:

- **Optimización: Adam ($\beta_1 = 0.5$, $\beta_2 = 0.999$)**

  El optimizador Adam (Kingma and Ba, 2014) es ampliamente utilizado en el entrenamiento de GANs debido a su eficiencia computacional y bajo requerimiento de memoria. Los parámetros $\beta_1 = 0.5$ y $\beta_2 = 0.999$ se seleccionan para estabilizar el entrenamiento adversarial, permitiendo un balance adecuado entre el generador y el discriminador (Radford et al., 2016).

- **Tasa de aprendizaje:** 2×10^{-4}

  Esta tasa de aprendizaje es común en tareas de colorización usando GANs, ya que permite un avance progresivo sin provocar oscilaciones ni colapsos en la red. Típicamente, la misma tasa se aplica tanto al generador como al discriminador para mantener simetría durante el entrenamiento (Zhu et al., 2017).

2.4.4. Aplicación en colorización

Las GANs han demostrado ser particularmente eficaces en tareas de colorización de imágenes, debido a su capacidad para generar resultados visualmente realistas y coherentes. Mientras que las CNN tradicionales pueden producir colorizaciones técnicamente correctas pero con colores apagados o irreales, las GANs añaden un componente perceptual que simula el estilo y la diversidad del color en imágenes reales (Nazeri et al., 2018).

El discriminador incentiva al generador a crear resultados que no solo cumplan con criterios matemáticos (como pérdida L2), sino que además "engañen" al ojo humano, aportando

saturación, coherencia y riqueza cromática. Gracias a este enfoque, las GANs han permitido alcanzar nuevos niveles de calidad en la colorización automática, incluso en contextos complejos como retratos o escenas naturales.

2.5. Vision Transformers (ViT)

Los *Transformers* fueron originalmente diseñados para tareas de procesamiento del lenguaje natural (Vaswani et al., 2017), pero más recientemente han sido adaptados a la visión computacional mediante el modelo *Vision Transformer* (ViT) (Dosovitskiy et al., 2021). A diferencia de las CNN, que operan con filtros convolucionales, ViT trata una imagen como una secuencia de "tokens visuales".

Este proceso inicia con la división de la imagen en pequeños parches (por ejemplo, de 16×16 píxeles), los cuales se proyectan a un espacio vectorial y se ingresan al modelo como si fueran palabras en un texto. El modelo *Transformer* emplea mecanismos de atención (*self-attention*) para aprender relaciones entre los parches, lo que le permite capturar dependencias globales sin recurrir a convoluciones.

Una ventaja clave de ViT es su capacidad para modelar relaciones a largo plazo entre distintas regiones de la imagen, lo que resulta especialmente útil en tareas de colorización con contextos complejos o composiciones espaciales amplias. Además, al no depender de *kernels* fijos, ViT puede ser más flexible y escalable que las CNN tradicionales, aunque generalmente requiere mayores cantidades de datos y potencia computacional para lograr un buen desempeño (Touvron et al., 2021).

2.5.1. Formulación matemática básica

En su lugar, divide una imagen en parches de tamaño fijo $p \times p$, los aplana en vectores y les añade codificación posicional para capturar relaciones espaciales.

La representación inicial se define como:

$$z_0 = [x_{\text{class}}; x_{p_1}E; x_{p_2}E; \ldots; x_{p_N}E] + E_{\text{pos}} \tag{2.7}$$

donde E es una matriz de *embeddings* y E_{pos} representa los *embeddings* posicionales. Esta secuencia de vectores se procesa mediante capas de atención multi-cabezal, lo cual permite al modelo capturar dependencias globales entre regiones de la imagen.

Esta capacidad de modelar relaciones a largo plazo es particularmente útil en tareas como la colorización automática de imágenes, donde mantener la coherencia semántica entre diferentes regiones es esencial (Dosovitskiy et al., 2021).

2.5.2. Componentes

El modelo *Vision Transformer* (ViT) reestructura el procesamiento de imágenes al tratar los fragmentos de una imagen como una secuencia, similar al texto en NLP (*Natural Language Processing*) o Procesamiento del Lenguaje Natural. Para lograr esto, se compone de varios elementos fundamentales que se detallan a continuación:

- **División de la imagen en parches** 16×16

  En lugar de aplicar convoluciones, ViT divide la imagen en parches fijos, comúnmente de tamaño 16×16 píxeles. Cada parche se aplana en un vector unidimensional. Por ejemplo, una imagen de tamaño 224×224 generará 196 parches (14×14), cada uno con

768 valores si la imagen es RGB. Esta partición convierte la imagen en una secuencia de *tokens*, lo que habilita el uso de arquitecturas tipo *Transformer* (Dosovitskiy et al., 2021).

- ***Embeddings* lineales**

  Cada parche se transforma mediante una proyección lineal (*embedding*) que lo representa en un espacio vectorial de dimensión fija, típicamente de 768 dimensiones. Este paso convierte los parches en "*tokens visuales*" comparables a palabras en un modelo de lenguaje. La calidad del *embedding* influye directamente en la capacidad del modelo para capturar las características visuales del contenido (Touvron et al., 2021).

- **Codificación Posicional (*Positional Encoding*)**

  Como los *Transformers* no tienen una estructura espacial inherente (a diferencia de las CNN), es necesario añadir un codificador posicional para que el modelo conserve información sobre la ubicación relativa de cada parche. Este *encoding* se suma al *embedding* de cada *token*, y puede ser fijo (basado en funciones seno y coseno) o aprendido durante el entrenamiento (Vaswani et al., 2017).

- **Capas de atención multi-cabeza (*Multi-Head Attention*)**

  El corazón del *Transformer* es la atención multi-cabeza, que permite al modelo observar diferentes regiones de la imagen simultáneamente y aprender relaciones entre parches. Cada "cabeza" de atención opera sobre distintas proyecciones del mismo *input*, lo que favorece una representación rica y variada. Este mecanismo es fundamental para capturar dependencias globales y contextuales, cruciales en la tarea de colorización (Dosovitskiy et al., 2021).

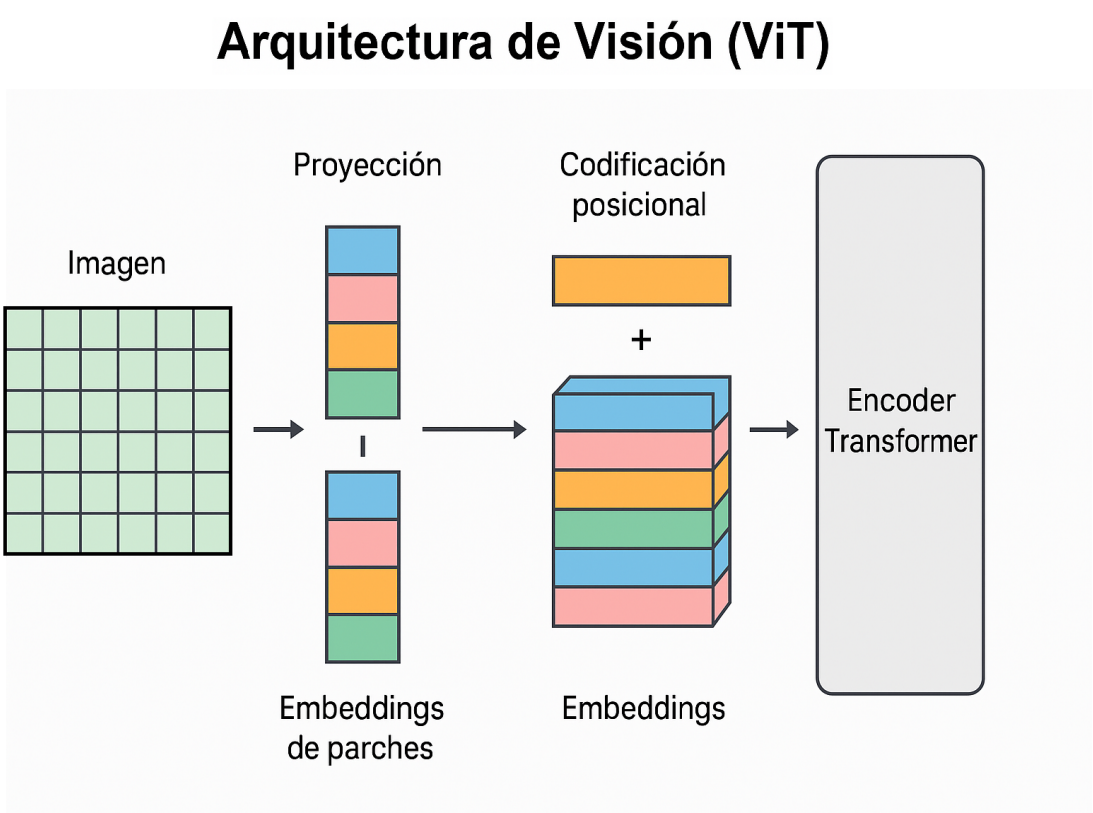


Figura 2.6: Arquitectura del *Vision Transformer* (ViT). La imagen se divide en parches, los cuales se proyectan a *embeddings*, se les suma codificación posicional y se ingresan a un *encoder* tipo *Transformer*.

La Figura 2.6 muestra el flujo completo de procesamiento de una imagen por un ViT. A la izquierda, se observa la segmentación de la imagen en parches; en el centro, cada parche es transformado en un vector y combinado con su codificación posicional; finalmente, todos los vectores se procesan en paralelo a través de múltiples capas del *encoder Transformer*. Este enfoque elimina la necesidad de convoluciones, enfocándose en relaciones globales entre regiones.

2.5.3. Hiperparámetros comunes

La arquitectura ViT incluye varios hiperparámetros clave que determinan su capacidad de representación y eficiencia. A continuación, se describen los más comunes:

- **Tamaño de parche:** 16×16

  Este tamaño de parche ha demostrado ser un compromiso efectivo entre granularidad espacial y eficiencia computacional. Parches más pequeños aumentan la resolución de la entrada pero también el costo computacional, mientras que parches más grandes reducen detalles locales. El parche de 16×16 ha sido ampliamente validado en tareas de clasificación y colorización (Dosovitskiy et al., 2021).

- **Dimensión del *embedding*:** 768

  La dimensión del espacio vectorial en el que se representa cada parche es fundamental para la capacidad del modelo. Un tamaño de 768 ofrece un equilibrio entre expresividad y consumo de memoria, permitiendo al ViT aprender características visuales complejas sin sobreajustarse (Touvron et al., 2021).

- **Número de cabezas de atención:** 12

  En la atención multi-cabeza, se usan 12 cabezas en paralelo, lo que permite al modelo observar diferentes aspectos del contexto de cada *token*. Esto mejora la habilidad para modelar relaciones espaciales y semánticas en imágenes, esenciales en la generación de color coherente (Vaswani et al., 2017).

- **Capas del *Transformer*:** 12

  El número de bloques *encoder* en el *Transformer* afecta la profundidad del modelo. ViT-base usa 12 capas, lo que permite una representación jerárquica suficiente sin incurrir en los altos costos computacionales de modelos más profundos como ViT-Large (Touvron et al., 2021). Estas capas incluyen operaciones de atención, normalización y perceptrones multicapa (*MLP*).

2.5.4. Aplicación en colorización

La aplicación de *Vision Transformers* en colorización de imágenes ha mostrado resultados prometedores, especialmente en la generación de colores coherentes y perceptualmente agradables en regiones amplias y con estructuras complejas.

A diferencia de las CNN tradicionales, que tienen un campo receptivo local, los ViT pueden capturar dependencias globales entre diferentes regiones de la imagen desde las primeras capas, gracias a su mecanismo de atención. Esta capacidad es especialmente útil en colorización, donde la coherencia semántica (por ejemplo, identificar que varios parches pertenecen al cielo o la ropa) es crucial para aplicar una colorización consistente.

Modelos como *Palette* (Saharia et al., 2022), un *Transformer* entrenado para colorización y otras tareas de reconstrucción, han superado en diversas métricas perceptuales a métodos basados en GANs o CNNs, generando resultados más vibrantes y realistas. Además, los ViTs son más robustos ante entradas ruidosas o degradadas, lo que los hace adecuados para colorización de imágenes históricas o de baja calidad.

2.6. Colorización base

La colorización base en este trabajo se entrenará con un modelo inicial de CNNs estándar, en el cual los valores cromáticos se predicen a partir de las características de luminancia de la imagen. Este enfoque clásico ha sido ampliamente documentado en estudios previos (Zhang et al., 2016; Larsson et al., 2016) y sirve como punto de referencia para evaluar la mejora que se logra con la integración de GANs y *Vision Transformers*.

Como punto de partida, se entrenará una CNN clásica que predice los canales a, b del espacio Lab. Las imágenes se convertirán a escala de grises (manteniendo el canal L), y se utilizarán

como entrada.

- Utiliza CNNs para estimar los valores de color de cada píxel a partir de características de textura y forma.
- Se basa en la representación en el espacio de color Lab, donde la luminancia (L) se mantiene y los canales de color (a, b) se predicen.
- No tiene un mecanismo de discriminación de realismo como las GANs, lo que puede generar colores poco realistas o sin contexto global.
- Se entrenará con imágenes a color convertidas a escala de grises antes de su procesamiento, asegurando que el modelo aprenda a predecir los colores de manera autónoma.

Además, se considerarán distintas funciones de pérdida para optimizar la calidad del color:

- ***Mean Squared Error*** **(MSE)**:

  La pérdida de error cuadrático medio mide la diferencia promedio entre los valores reales y los valores predichos por el modelo. Es útil cuando se desea minimizar directamente la distancia entre los colores reales y generados, evaluando cada píxel individualmente.

$$\mathcal{L}_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.8}$$

  donde:

  - y_i representa el valor real del color en el píxel i,

- $\hat{y}_i$ es el valor predicho por el modelo en ese mismo píxel,
- n es el número total de píxeles.

Esta pérdida es sencilla y efectiva, pero suele generar imágenes borrosas, ya que penaliza todas las desviaciones del valor promedio, incluso si son perceptualmente aceptables (Zhang et al., 2018).

- ***Perceptual Loss***:

  La pérdida perceptual no compara directamente los valores de píxeles, sino que evalúa la diferencia entre las representaciones de alto nivel extraídas de una red convolucional preentrenada, como VGG-16 (Simonyan and Zisserman, 2015). Estas representaciones están correlacionadas con la percepción humana de similitud visual.

$$\mathcal{L}_{perc} = \sum_{l} \|\phi_l(y) - \phi_l(\hat{y})\|_2^2 \tag{2.9}$$

  donde:

  - $\phi_l(\cdot)$ representa la activación en la capa l de una red preentrenada (por ejemplo, VGG-16),
  - y es la imagen real,
  - $\hat{y}$ es la imagen generada.

  Esta función favorece la generación de texturas y detalles realistas, y ha sido ampliamente utilizada en tareas de reconstrucción de imágenes (Johnson et al., 2016).

- ***Loss adversarial* en GANs**:

  Optimización mediante retroalimentación del discriminador. En redes generativas adversariales (GANs), el generador es optimizado no solo para minimizar una pérdida como MSE o perceptual, sino también para engañar al discriminador, quien intenta

distinguir entre imágenes reales y generadas. Esta retroalimentación adversarial fomenta la creación de imágenes más fotorrealistas.

La función de pérdida adversarial para el generador se define como:

$$L_{\text{adv}} = -\log(D(G(x))) \tag{2.10}$$

donde:

- $G(x)$ es la imagen generada por el generador a partir de la entrada x,
- $D(\cdot)$ es la probabilidad que asigna el discriminador a que una imagen sea real.

Por su parte, el discriminador se entrena con la siguiente pérdida:

$$L_D = -\left[\log(D(y)) + \log(1 - D(G(x)))\right] \tag{2.11}$$

Estas pérdidas se combinan frecuentemente con MSE o Perceptual Loss para equilibrar precisión y realismo. Este enfoque híbrido es común en colorización con GANs, como en los trabajos de Zhang et al. (2016).

2.7. Evaluación de la mejora

Las imágenes de prueba se seleccionarán de un conjunto independiente de datos que no haya sido visto por el modelo durante el entrenamiento, asegurando una evaluación realista de su capacidad de generalización. Para medir la mejora con respecto a la colorización base, se utilizarán las siguientes métricas:

- PSNR (*Peak Signal-to-Noise Ratio*): Evalúa la diferencia entre la imagen generada y

la imagen de referencia. Un PSNR más alto indica menor error en la reconstrucción del color.

- SSIM (*Structural Similarity Index*): Mide la similitud estructural entre la imagen generada y la imagen de referencia, evaluando el contraste, brillo y textura.
- Evaluación perceptual con *Perceptual Loss*: Se compararán características de alto nivel en la imagen generada con una imagen real, permitiendo ajustar la calidad cromática y semántica de la colorización.
- Evaluación subjetiva: Se realizarán pruebas con observadores humanos para medir la percepción de realismo y coherencia del color.

La expectativa es que la combinación de *Vision Transformers* y GANs genere imágenes con colores más naturales y consistentes en comparación con las CNNs tradicionales, asegurando una mejora tanto en las métricas cuantitativas como en la percepción subjetiva de la calidad de la imagen.

2.8. Tendencias y retos actuales

El campo de la visión por computadora ha experimentado un cambio paradigmático con la introducción y maduración del aprendizaje profundo (*Deep Learning*, DL), el cual ha desplazado el enfoque tradicional basado en ingeniería manual de características hacia uno en el que las redes neuronales profundas son capaces de aprender representaciones jerárquicas directamente desde imágenes crudas. Este avance ha sido fundamental para tareas como clasificación, segmentación, detección de objetos, reconstrucción y, de manera destacada, la colorización automática de imágenes en escala de grises.

No obstante, a pesar del éxito de las arquitecturas modernas, el diseño efectivo de redes

profundas sigue siendo un proceso complejo y altamente especializado. Se requieren conocimientos avanzados en optimización, diseño de arquitectura y experiencia empírica para desarrollar modelos eficientes y precisos (Goodfellow et al., 2014). A medida que los modelos se vuelven más sofisticados—como los híbridos que combinan redes neuronales convolucionales (CNN), redes generativas adversariales (GAN) y *Vision Transformers* (ViT), también aumentan los desafíos asociados a su entrenamiento, interpretación y despliegue.

2.8.1. Evolución arquitectónica y necesidad de modelos eficientes

La evolución de las CNN ha demostrado que una mayor profundidad y complejidad arquitectónica puede mejorar significativamente la capacidad de representación de los modelos. Sin embargo, esto ha llevado a modelos cada vez más pesados y demandantes de recursos computacionales, lo que limita su aplicación en entornos con baja capacidad de procesamiento o en aplicaciones en tiempo real (He et al., 2016; Tan and Le, 2019). Por ello, ha surgido una tendencia hacia el diseño de arquitecturas más ligeras y eficientes, como MobileNet, EfficientNet y variantes optimizadas de Transformers, que mantienen una alta precisión con menor tamaño y complejidad (Howard et al., 2019).

El reto es aún mayor cuando se aplican estas arquitecturas en tareas generativas como la colorización. En este contexto, el modelo debe no solo aprender una representación semántica robusta, sino también generar resultados visualmente plausibles, coherentes y naturales. La combinación de CNNs para la extracción local de características, GANs para la generación realista del color y ViTs para capturar relaciones espaciales globales ha demostrado ser prometedora, aunque también plantea retos de integración, entrenamiento y evaluación (Isola et al., 2017; Dosovitskiy et al., 2021).

2.8.2. Limitaciones del aprendizaje profundo y el problema de los datos

Uno de los problemas fundamentales del DL es la necesidad de grandes volúmenes de datos etiquetados para un entrenamiento efectivo. En la práctica, como en la colorización de imágenes, disponer de millones de ejemplos perfectamente anotados (pares de imágenes en escala de grises y color reales) no siempre es factible, y la adquisición de estos datos puede ser costosa en tiempo y recursos humanos (Zhang et al., 2016). Para mitigar este problema, se han explorado estrategias como el aumento de datos, el uso de modelos preentrenados y, más recientemente, el aprendizaje auto-supervisado, que representa una de las tendencias más prometedoras cuando el etiquetado es escaso o costoso (Chen et al., 2020).

Además, los modelos generativos como los GANs presentan dificultades inherentes: su entrenamiento es inestable debido a la dinámica entre generador y discriminador, pueden sufrir de colapso del modo y requieren cuidadosa sintonización de hiperparámetros y funciones de pérdida (Arjovsky et al., 2017). Cuando se integran componentes como ViTs, originalmente diseñados para clasificación o NLP, se requiere rediseñar estrategias arquitectónicas y de entrenamiento para lograr una integración efectiva en pipelines generativos.

2.8.3. Retos semánticos y perceptuales en colorización

Desde un punto de vista perceptual, la colorización no es una tarea trivial. Una misma imagen en escala de grises puede tener múltiples coloraciones plausibles. Los modelos deben inferir no solo la estructura y forma de los objetos, sino también realizar inferencias semánticas complejas para decidir, por ejemplo, si una fruta es una manzana roja, verde o amarilla; o si un cielo debe teñirse de gris azulado o naranja al atardecer (Iizuka et al., 2016). Estas decisiones requieren que el modelo integre elementos contextuales de alto nivel, donde los

ViTs ofrecen una ventaja por su capacidad de capturar relaciones de largo alcance dentro de la imagen.

Un reto adicional es la evaluación objetiva de los resultados. Métricas tradicionales como PSNR o SSIM, aunque útiles, no siempre se correlacionan con la percepción humana de calidad del color. Esto ha motivado la exploración de métricas perceptuales más robustas y evaluaciones subjetivas mediante estudios con usuarios (Zhang et al., 2018).

2.8.4. Sostenibilidad, escalabilidad y procesamiento en tiempo real

El entrenamiento de modelos grandes, en particular aquellos que involucran GANs y ViTs, es intensivo en términos de energía y requiere *hardware* especializado como GPUs (*Graphics Processing Units*) o TPUs (*Tensor Processing Units*). Esto ha generado preocupaciones sobre el impacto ambiental del desarrollo de modelos de DL a gran escala (Strubell et al., 2019). Para aplicaciones móviles, interactivas o en sistemas embebidos, es esencial optimizar los modelos no solo en precisión, sino también en velocidad y eficiencia energética.

Para ello, se han desarrollado estrategias como la compresión de modelos, la poda de parámetros, la cuantización y la destilación de conocimiento como alternativas viables para reducir el tamaño sin degradar significativamente el rendimiento (Cheng et al., 2018). Estas estrategias son también aplicables a sistemas de colorización, especialmente si se planea su implementación en contextos industriales o comerciales con restricciones técnicas.

2.9. Aporte de la investigación

El presente trabajo busca mejorar la colorización de imágenes mediante la integración de *Vision Transformers* y GANs, lo que podría:

- Reducir la incoherencia cromática presente en modelos basados únicamente en CNNs.
- Mejorar la interpretación contextual de los colores en grandes regiones de la imagen.
- Aumentar la precisión en la representación de detalles complejos.

A diferencia de estudios previos que han explorado CNNs o GANs de manera independiente, este proyecto propone un modelo híbrido *Vision Transformers* + GANs, que integra la capacidad de aprendizaje contextual de los *Transformers* con la generación realista de los GANs. Con este enfoque, se espera obtener imágenes con una mayor coherencia cromática a nivel global, evitando errores frecuentes en transiciones de color y detalles finos, lo que representa un avance, abriendo nuevas posibilidades en la automatización del colorizado de imágenes.

Capítulo 3

Propuesta Metodológica

Esta investigación se llevó a cabo con el propósito de diseñar y entrenar un sistema automático de colorización de imágenes en escala de grises, utilizando un modelo híbrido que integra redes neuronales convolucionales (CNN), redes generativas adversariales (GAN), *Vision Transformers* (ViT) y técnicas de pérdida perceptual extraída de un modelo preentrenado (VGG-16). Esta selección metodológica busca aprovechar las ventajas particulares de cada arquitectura para mejorar la fidelidad, coherencia semántica y calidad visual de las imágenes colorizadas

La Figura 3.1 muestra de forma esquemática la arquitectura general del modelo híbrido propuesto, en donde se ilustran los flujos de datos desde la imagen en escala de grises hasta la reconstrucción final colorizada. Se observa cómo las CNN actúan como extractores de características iniciales, el *Vision Transformer* interpreta el contenido semántico global, y el GAN genera imágenes con color realista, mientras que la pérdida perceptual actúa como guía de alto nivel para el entrenamiento.

El proceso se estructura en cuatro fases principales:

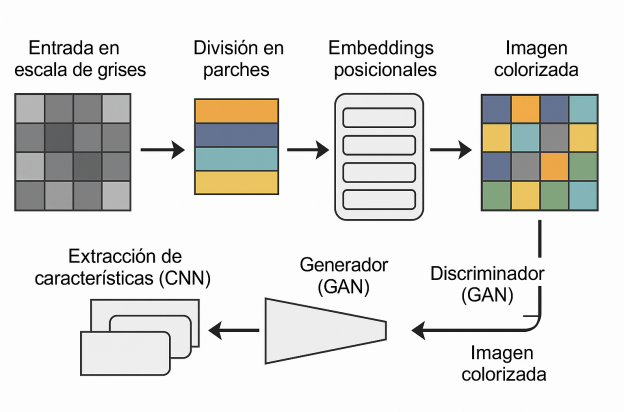


Figura 3.1: Modelo híbrido

1. Preprocesamiento de datos.
2. Diseño de implementación del modelo híbrido.
3. Entrenamiento del modelo y evaluación cuantitativa.
4. Pruebas y visualización de resultados.

3.1. Preprocesamiento de datos

Se utilizó el conjunto de datos Tiny ImageNet, una versión reducida del ImageNet, que contiene 100,000 imágenes a color de 64×64 píxeles, divididas en 200 clases, asi como también que es el datasets final, el ImageNet completo.

Las imágenes fueron convertidas del espacio de color RGB al espacio CIELAB Zhang et al. (2016), donde:

- L representa la luminancia (intensidad de luz).
- a y b codifican la información cromática (verde-rojo y azul-amarillo).

Las imágenes son normalizadas del siguiente modo:

$$L_{\text{norm}} = \frac{L}{100}, \quad A_{\text{norm}} = \frac{A}{127}, \quad B_{\text{norm}} = \frac{B}{127} \tag{3.1}$$

Estas normalizaciones permiten que los valores estén dentro de un rango estable de $[-1, 1]$ para facilitar el entrenamiento del modelo, como recomiendan Ioffe y Szegedy (2015).

Ademas se convierte la imagen de entrada del espacio RGB al espacio CIELAB, que separa la luminancia de los componentes cromáticos. Esto facilita que el modelo aprenda a predecir únicamente los colores a partir de la luminancia (L^*), manteniendo la estructura de la imagen.

3.1.1. Conversión de RGB a XYZ

La conversión se realiza primero desde RGB hacia el espacio de color CIEXYZ, utilizando la matriz de transformación definida en la Ecuación (3.2):

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.412453 & 0.357580 & 0.180423 \\ 0.212671 & 0.715160 & 0.072169 \\ 0.019334 & 0.119193 & 0.950227 \end{bmatrix} \cdot \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{3.2}$$

Donde R, G, B son los valores de los canales rojo, verde y azul normalizados entre $[0, 1]$, y X, Y, Z representan los componentes en el espacio CIEXYZ. Esta transformación es un paso intermedio necesario para obtener una representación perceptualmente uniforme del color.

3.1.2. Conversión de XYZ a CIELAB

Una vez en el espacio CIEXYZ, los valores se convierten al espacio CIELAB utilizando las siguientes fórmulas, presentadas en la ecuación (3.3).

$$\begin{aligned} L &= 116f\left(\frac{Y}{Y_n}\right) - 16 \\ a &= 500\left[f\left(\frac{X}{X_n}\right) - f\left(\frac{Y}{Y_n}\right)\right] \\ b &= 200\left[f\left(\frac{Y}{Y_n}\right) - f\left(\frac{Z}{Z_n}\right)\right] \end{aligned} \tag{3.3}$$

Donde:

- L: valor de luminancia perceptual, es decir, la cantidad de luz que una persona percibe. Tiene un rango de valores aproximado entre 0 (negro) y 100 (blanco), y es responsable de definir el brillo de la imagen. Es el canal que se conserva como entrada en el proceso de colorización automática, ya que mantiene la estructura de la imagen original en escala de grises.
- a: componente verde-rojo. Valores negativos indican tonos verdosos, mientras que valores positivos indican tonos rojizos. Este canal, junto con el canal b, es el que el modelo debe aprender a predecir durante el entrenamiento, basándose únicamente en la luminancia L.
- b: componente azul-amarillo. Valores negativos indican tonalidades azuladas, y valores positivos indican tonos amarillentos. Al igual que a, este componente es generado por el modelo a partir de la luminancia como parte del proceso de colorización.
- X_n, Y_n, Z_n: Son los valores de referencia del blanco bajo condiciones estándar de iluminación. Para lograr una conversión adecuada y consistente, se deben usar valores

específicos del blanco de referencia (por ejemplo, para D65: $X_n = 95.047$, $Y_n = 100.0$, $Z_n = 108.883$).

La función $f(t)$ que aparece en las ecuaciones anteriores se define como una corrección no lineal y está descrita en la Ecuación (3.4):

$$f(t) = \begin{cases} t^{1/3}, & \text{si } t > 0.008856 \\ 7.787t + \frac{16}{116}, & \text{si } t \leq 0.008856 \end{cases} \tag{3.4}$$

Esta función garantiza una representación perceptual más precisa de las diferencias de color, especialmente para valores bajos de luminancia.

3.2. Descripción detallada del sistema propuesto

El sistema propuesto se basa en una arquitectura híbrida compuesta por tres módulos principales:

- **Módulo de extracción de características (CNN):** Se encarga de capturar patrones locales y espaciales de la imagen en escala de grises. Este módulo transforma la imagen en un conjunto de características representativas de bajo y medio nivel.
- **Módulo de atención global (ViT):** Se encarga de modelar relaciones de largo alcance entre regiones distantes de la imagen. Utiliza mecanismos de autoatención para capturar la estructura semántica global, lo cual contribuye a una colorización más coherente.

- **Módulo generador y discriminador (GAN):** El generador toma las características extraídas y produce los canales cromáticos (a, b) en el espacio CIELAB, los cuales se combinan con el canal de luminancia (L) para reconstruir la imagen en color. El discriminador evalúa la autenticidad de las imágenes generadas frente a las imágenes reales, mejorando la capacidad del generador.

3.3. Arquitectura del modelo híbrido

El modelo completo fue compuesto por dos módulos principales: Generador (G) y Discriminador (D). Se diseñó un modelo con una arquitectura tipo UNet para el generador, con codificador-decodificador y conexiones tipo *skip* entre capas correspondientes. Este generador transforma imágenes en escala de grises (canal L del espacio de color CIELAB) en los canales a y b que representan la crominancia. La salida del generador es posteriormente evaluada por un discriminador tipo PatchGAN, que intenta distinguir entre imágenes colorizadas reales y generadas.

3.3.1. Generador

El generador fue construido a partir de una arquitectura UNet mejorada, que utiliza bloques de convolución y transconvolución con normalización por instancia (*InstanceNorm*) y funciones de activación no lineales (LeakyReLU y ReLU), permitiendo una reconstrucción efectiva de las características a través de conexiones de salto (*skip connections*) (Ronneberger et al., 2015). El generador recibió como entrada una imagen en escala de grises y generó dos canales de color $\hat{a}$ y $\hat{b}$. Su arquitectura combinó capas convolucionales con normalización de instancia y activaciones LeakyReLU, seguidas por capas transpuestas (*upsampling*). Las conexiones de salto permiten preservar detalles espaciales finos.

3.3.2. Discriminador

Clasifica regiones locales de la imagen en lugar de toda la imagen, permitiendo una retroalimentación de mayor resolución. El discriminador se implementó siguiendo el paradigma PatchGAN (Isola et al., 2017), lo que permite juzgar la calidad de pequeñas regiones de la imagen generada en lugar de evaluar la imagen completa, promoviendo la generación de detalles coherentes a nivel local.

3.3.3. Fórmulas del modelo

- **Red Neuronal Convolucional (CNN) — Extracción de características locales** Las redes neuronales convolucionales (CNN) están diseñadas específicamente para procesar datos con una estructura en forma de rejilla, como las imágenes. Estas redes aplican filtros (*kernels*) que se deslizan sobre la imagen para extraer características espaciales locales, como bordes, texturas o patrones complejos. Cada filtro aprende a detectar un tipo específico de característica a través del proceso de entrenamiento (LeCun et al., 2015).

  El principio básico de operación de una CNN puede entenderse como una extensión del modelo de neurona artificial descrito previamente en la Ecuación (2.1), donde en lugar de realizar una suma ponderada simple sobre un vector de entrada, se realiza una operación de convolución sobre regiones locales de una imagen. Posteriormente, el resultado es transformado mediante una función de activación no lineal, como ReLU (ya definida en la Ecuación 2.2).

  Este mecanismo permite que las CNN capten información jerárquica y espacial de manera muy eficiente, siendo fundamentales en tareas como clasificación de imágenes, detección de objetos y, particularmente en este caso, colorización automática, donde los detalles locales y la coherencia estructural son críticos para la calidad visual del

resultado.

- **Vision Transformer — Atención global** Los Transformers capturan relaciones entre píxeles distantes usando mecanismos de atención.

  Atención escalada (*Self-Attention*)

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.5}$$

  Donde:

  - Q: matriz de consultas ($n \times d_k$)
  - K: matriz de claves ($n \times d_k$)
  - V: matriz de valores ($n \times d_k$)
  - d_k: dimensión de las claves
  - n: número de parches de entrada
  - *softmax*: convierte puntajes en probabilidades

- **Discriminador (GAN):**

  Se encarga de diferenciar entre imágenes reales (coloreadas) y generadas (Goodfellow et al., 2014).

$$D(x) \in [0, 1] \tag{3.6}$$

  - $D(x)$: probabilidad estimada de que la imagen x sea real.

  Función de pérdida del discriminador:

$$\mathcal{L}_D = -\mathbb{E}_{x\sim p_{\text{real}}}[\log D(x)] - \mathbb{E}_{\hat{x}\sim p_{\text{fake}}}[\log(1 - D(\hat{x}))] \tag{3.7}$$

Donde:

- $\mathbb{E}$: esperanza matemática
- x: imagen real, $\hat{x}$: imagen generada
- p_{real}, p_{fake}: distribuciones reales y generadas

- **Generador:**

Función de pérdida híbrida (entrenamiento del generador)

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{MSE}} + \lambda_2\mathcal{L}_{\text{adv}} + \lambda_3\mathcal{L}_{\text{percep}} + \lambda_4\mathcal{L}_{\text{TV}} \tag{3.8}$$

Donde $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ son coeficientes de ponderación.

Esta combinación permite balancear la precisión local, la percepción visual y el realismo global en las imágenes generadas (Zhang et al., 2018).

a) Error cuadrático medio (MSE)

Definida en la Ecuación (2.8)

Donde:

- N: número de píxeles
- y_i: valor real del canal a^* o b^*
- $\hat{y}_i$: valor predicho por el generador

Esta función mide la distancia promedio entre los valores predichos y reales de los canales de color (Zhou et al., 2016).

b) Pérdida adversarial (GAN)

$$\mathcal{L}_{\text{adv}} = -\log(D(G(x))) \tag{3.9}$$

- $D(G(x))$: probabilidad de que la imagen generada sea real según el discriminador.
- $G(x)$: salida del generador a partir de la imagen en escala de grises.

Esta función incentiva al generador a producir imágenes que el discriminador no pueda distinguir de las reales (Goodfellow et al., 2014).

c) Pérdida perceptual (usando VGG16)

$$\mathcal{L}_{\text{percep}} = \sum_{l} \frac{1}{H_l W_l} \|\phi_l(y) - \phi_l(\hat{y})\|_2^2 \tag{3.10}$$

Donde:

- $\phi_l(y)$: activación de capa l para la imagen real
- $\phi_l(\hat{y})$: activación para la imagen generada
- H_l, W_l: altura y ancho de mapas de activación

Esta pérdida busca preservar la estructura visual global en lugar de solo minimizar diferencias píxel a píxel (Johnson et al., 2016).

d) Pérdida de variación total (TV)

$$\mathcal{L}_{\text{TV}} = \sum_{i,j} \left((\hat{y}_{i,j+1} - \hat{y}_{i,j})^2 + (\hat{y}_{i+1,j} - \hat{y}_{i,j})^2 \right) \tag{3.11}$$

- $\hat{y}_{i,j}$: valor del píxel generado en la posición (i, j).
- Penaliza grandes diferencias entre píxeles vecinos para suavizar el color.

- **Métricas de evaluación**

a) PSNR (*Peak Signal-to-Noise Ratio*)

El PSNR es una métrica ampliamente utilizada para evaluar la calidad de una imagen reconstruida o generada en comparación con una imagen original de referencia. Se expresa en decibelios (dB) y mide la proporción entre la señal máxima posible (el valor máximo del píxel) y el error introducido por la reconstrucción (Wang et al., 2004).

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{MAX^2}{\text{MSE}} \right) \tag{3.12}$$

Donde:

- MAX es el valor máximo posible de un píxel (255 para imágenes de 8 bits por canal).
- MSE (*Mean Squared Error*) es el error cuadrático medio entre la imagen original y la generada

b) SSIM (*Structural Similarity Index*)

El SSIM evalúa la calidad de una imagen generada comparando su estructura, luminancia y contraste con respecto a una imagen de referencia. A diferencia del PSNR, el SSIM intenta imitar la percepción humana de calidad visual (Wang et al., 2004).

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{3.13}$$

Donde:

- μ_x, μ_y: medias locales de las imágenes x y y.
- σ_x^2, σ_y^2: varianzas locales.
- σ_{xy}: covarianza local entre x y y.
- C_1, C_2: constantes pequeñas para evitar divisiones por cero (comúnmente, $C_1 = (0.01 \cdot L)^2$ y $C_2 = (0.03 \cdot L)^2$, con $L = 255$ para imágenes de 8 bits).

3.4. Diagrama del sistema

En la Figura 3.2 se muestra el diagrama a bloques del sistema propuesto. Este representa el flujo de datos desde la entrada de la imagen en escala de grises hasta la generación final de la imagen colorizada.

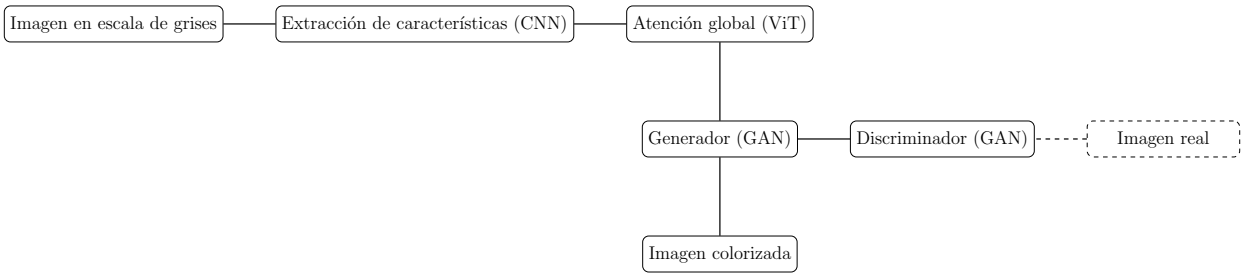


Figura 3.2: Diagrama de bloques del sistema propuesto

3.4.1. Procedimiento de entrenamiento

El modelo híbrido fue entrenado durante 100 épocas utilizando un tamaño de lote de 32 imágenes. En cada iteración del proceso de entrenamiento se llevaron a cabo las siguientes etapas:

1. **Carga de datos:** Se seleccionó un lote de entrada X (imágenes en escala de grises) y sus correspondientes etiquetas Y (versiones colorizadas reales).

2. **Generación de predicciones:** El generador G produjo una estimación colorizada $\hat{Y} = G(X)$.

3. **Entrenamiento del discriminador:** Se entrenó el discriminador D para distinguir entre las imágenes reales Y y las generadas $\hat{Y}$.

4. **Cálculo de pérdida y actualización del generador:** Se calculó la función de pérdida total, compuesta por una combinación de pérdida adversarial, pérdida perceptual

y pérdida basada en similitud estructural. A partir de esta pérdida, se retropropagó el error y se actualizaron los parámetros del generador para mejorar la calidad y coherencia de las imágenes generadas.

El optimizador utilizado fue Adam con tasa de aprendizaje 2×10^{-4}, $\beta_1 = 0.5$, $\beta_2 = 0.999$.

3.5. Herramientas y recursos tecnológicos

Para llevar a cabo el desarrollo del sistema se emplearán las siguientes herramientas:

- **Lenguaje de programación:** Python 3.6 (por compatibilidad con servidores PPC64).
- **Framework de Deep Learning:** PyTorch 1.3.1.
- **Hardware:** Sistema IBM con 4 GPUs NVIDIA Tesla V100 (Compute Capability 7.0) y arquitectura PPC64LE.
- **Dataset:** *Tiny ImageNet*, preprocesado al espacio CIELAB.

3.5.1. Conjuntos de datos utilizados

Para el desarrollo y evaluación del sistema de colorización automática de imágenes se utilizaron dos conjuntos de datos relevantes dentro del ámbito del aprendizaje profundo: Tiny ImageNet y ImageNet. Ambos forman parte de la familia de datasets derivados del proyecto ImageNet, ampliamente reconocido por su papel fundamental en el avance de las arquitecturas de visión por computadora.

- **Tiny ImageNet**: versión reducida de ImageNet del conjunto de datos ImageNet, diseñada específicamente para facilitar la experimentación con modelos que requieren me-

nos recursos computacionales. Esta base de datos contiene aproximadamente 100,000 imágenes a color con una resolución de 64×64 píxeles, distribuidas en 200 clases diferentes. Cada clase cuenta con 500 imágenes para entrenamiento, 50 para validación y 50 para prueba.

- **ImageNet (Zhang et al., 2016)**: Es una de las bases de datos más grandes y completas utilizadas en la investigación de visión por computadora. Fue creada como parte del proyecto ImageNet y contiene más de 14 millones de imágenes etiquetadas manualmente en más de 20,000 categorías de objetos. Su versión más conocida y utilizada es la que se empleó en el desafío ImageNet *Large Scale Visual Recognition Challenge* (ILSVRC), que incluye imágenes de resolución más alta (generalmente mayores a 224×224 píxeles) organizadas en 1,000 clases.

3.6. Criterios de evaluación del sistema

La calidad de la colorización generada por el modelo propuesto será evaluada mediante dos métricas cuantitativas ampliamente aceptadas en el ámbito de la visión por computadora:

- **PSNR (Peak Signal-to-Noise Ratio):** Esta métrica cuantifica la diferencia entre la imagen generada y la imagen original a nivel de píxeles. Es particularmente útil para medir la fidelidad y precisión de la reconstrucción, ya que penaliza fuertemente los errores de color en regiones específicas. Una mayor puntuación de PSNR indica menor distorsión y, por lo tanto, una mejor aproximación a la imagen real.
- **SSIM (Structural Similarity Index):** A diferencia del PSNR, esta métrica evalúa la percepción visual y la coherencia estructural entre dos imágenes, considerando factores como luminancia, contraste y estructura. El SSIM es útil para capturar diferencias que

son relevantes desde el punto de vista humano, evaluando no solo los errores puntuales, sino también cómo se percibe globalmente la imagen.

La combinación de estas métricas permite una evaluación integral del sistema, ya que PSNR proporciona una medida objetiva y matemática del error, mientras que SSIM incorpora aspectos de percepción visual. Su uso conjunto está ampliamente respaldado en la literatura (Wang et al., 2004) y permite comparar de manera más justa el desempeño del modelo híbrido propuesto (CNN + GAN + ViT) frente a enfoques tradicionales basados exclusivamente en CNN o GAN, destacando sus ventajas en términos de precisión cromática y coherencia contextual.

Capítulo 4

Experimentación y Resultados

Este capítulo expone el proceso experimental diseñado para evaluar el desempeño de los modelos de colorización automática propuestos. Se realiza un análisis comparativo entre dos arquitecturas principales como paso inicial, con el propósito de introducir de manera ordenada los modelos desarrollados. Esta estrategia permite una incorporación progresiva de los componentes arquitectónicos —CNN, GAN y ViT— con el fin de observar su impacto individual y combinado en la calidad de la colorización. Posteriormente, se integran los tres modelos en una arquitectura híbrida, lo cual proporciona una visión más completa del comportamiento del sistema y aporta evidencia sólida para respaldar los objetivos de esta investigación.

- **Modelo Base (CNN+GAN)**: Arquitectura convencional basada en redes generativas adversarias con generador U-Net y discriminador PatchGAN.
- **Modelo Híbrido (CNN+ViT+GAN)**: Extensión del base que incorpora bloques de Transformers para capturar dependencias globales.

El análisis incluye configuración experimental, protocolo de entrenamiento, métricas cuanti-

tativas y evaluación cualitativa de resultados. Todos los experimentos se realizaron utilizando el dataset *ImageNet* bajo idénticas condiciones de *hardware* (GPU NVIDIA V100 32GB) para garantizar la comparabilidad.

4.1. Arquitectura del modelo CNN + GAN

El modelo implementado sigue un esquema Generativo Adversarial (GAN) con las siguientes características:

Generador (U-Net mejorado)

***Encoder*:**

- Tres capas convolucionales con normalización por instancias y activación LeakyReLU.
- Incremento progresivo de canales: $64 \rightarrow 128 \rightarrow 256$.
- Inclusión de tres bloques residuales para preservar la información espacial.

***Bottleneck*:**

- Capas convolucionales dilatadas con `dilation=2`, lo que incrementa el campo receptivo sin reducir la resolución espacial.

***Decoder*:**

- Dos capas convolucionales con *upsampling* bilineal para reconstruir la imagen a su resolución original.
- Función de activación final `Tanh()` para restringir la salida al rango $[-1, 1]$.

Discriminador (PatchGAN)

- Cinco capas convolucionales con normalización por instancias y activación LeakyReLU.
- Aplicación de `Dropout` con $p = 0.2$ para reducir el sobreajuste.
- Salida de dimensión 30×30 que permite clasificar regiones locales de la imagen en lugar de evaluarla globalmente.

4.1.1. Estrategia de entrenamiento

Durante el entrenamiento del modelo híbrido de colorización, se emplearon diversos hiperparámetros cuidadosamente seleccionados para garantizar un equilibrio entre rendimiento, estabilidad y calidad visual de las imágenes generadas. La Tabla 4.1 resume estos valores clave.

Hiperparámetros

Tabla 4.1: Hiperparámetros utilizados en el entrenamiento

Parámetro	**Valor**	**Descripción**
Tamaño de *batch*	32	Balance entre uso de memoria y estabilidad del gradiente.
Dimensión de imagen	256×256	Resolución suficiente para detalles finos sin exceso computacional.
Learning Rate (G)	0.0002	Mayor que el del discriminador para compensar la pérdida L1.
Learning Rate (D)	0.0001	Valor conservador para evitar colapso del discriminador.
λ (L1 *Loss*)	50	Peso que regula el balance entre realismo y fidelidad al *ground truth*.
Optimizador	Adam	Momentum ($\beta_1 = 0.5$, $\beta_2 = 0.999$) para evitar oscilaciones.

4.1.2. Tabla Resumen de métricas clave

La Tabla 4.2 presenta un resumen de las métricas clave recolectadas durante las primeras cinco épocas del entrenamiento del modelo híbrido. En ella se observan valores correspondientes a la pérdida del generador (**Pérdida G**), la pérdida del discriminador (**Pérdida D**), así como la pérdida L1, que refleja la diferencia directa entre la imagen real y la generada. Además, se incluyen las métricas de calidad **PSNR** y **SSIM**, que miden la fidelidad visual de las imágenes colorizadas en relación con las originales. Finalmente, se reporta el tiempo promedio por época en horas.

Tabla 4.2: Resumen de métricas por época

Época	Pérdida G	Pérdida D	Pérdida L1	PSNR (dB)	SSIM	Tiempo/Época (h)
1	5.997	0.528	0.095	21.55	0.995	3.05
2	8.369	0.223	0.099	23.01	0.993	3.06
3	9.122	0.148	0.098	22.74	0.995	3.04
4	9.815	0.099	0.096	22.55	0.996	3.04
5	11.348	0.062	0.100	23.53	0.996	3.05

4.1.3. Interpretación del comportamiento del modelo

Durante las primeras cinco épocas de entrenamiento, se observó una dinámica particular en la evolución de las métricas que revela aspectos fundamentales del proceso de aprendizaje. La pérdida del generador mostró un incremento progresivo desde 5.997 en la primera época hasta 11.348 en la quinta, comportamiento que, aunque aparentemente contradictorio, es característico en arquitecturas GAN competitivas donde el generador debe realizar un esfuerzo cada vez mayor para producir salidas convincentes que puedan engañar a un discriminador que se está volviendo más sofisticado en su tarea de distinguir entre imágenes reales y generadas.

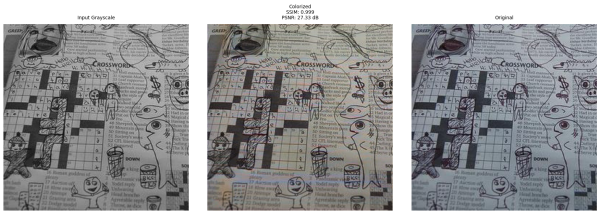
Por otro lado, la pérdida del discriminador presentó una disminución constante desde 0.528 hasta 0.062 en el mismo período. Esta reducción progresiva indica que, efectivamente, el discriminador va perdiendo capacidad para diferenciar las imágenes generadas de las reales, lo cual es un indicador positivo del aprendizaje del generador. Sin embargo, esta rápida convergencia del discriminador también podría sugerir la necesidad de ajustar el balance entre las tasas de aprendizaje de ambos componentes para mantener una competencia más equilibrada durante el entrenamiento.

Las métricas de calidad PSNR y SSIM mostraron valores altos desde las primeras épocas, con el SSIM manteniéndose consistentemente por encima de 0.99, lo que indica una excelente preservación de la estructura de las imágenes. El PSNR, que partió de 21.55 dB en la primera época, experimentó una mejora hasta alcanzar 23.53 dB en la quinta época, reflejando una reducción gradual en el error cuadrático medio entre las imágenes generadas y las reales. Este incremento en el PSNR, aunque moderado, es significativo considerando el corto período de entrenamiento analizado.

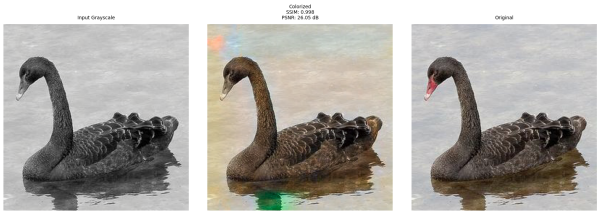
4.1.4. Evaluación cualitativa del desempeño

Al examinar visualmente las muestras generadas en diferentes etapas del entrenamiento, se aprecia una evolución notable en la calidad de las colorizaciones. En la primera época, las imágenes presentaban una paleta cromática limitada, dominada principalmente por tonos fríos como azules y verdes, con una saturación notablemente baja. Esta etapa inicial también mostraba artefactos visibles en regiones con texturas complejas.

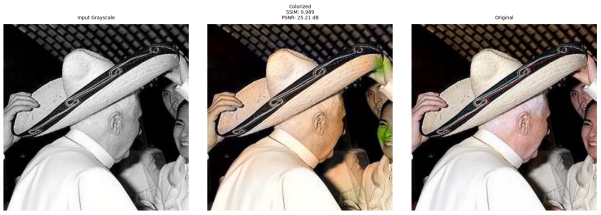
(a) Época 1



(b) Época 2



(c) Época 3



(d) Época 4

(e) Época 5

Figura 4.1: Evolución visual de la colorización generada por el modelo durante las cinco primeras épocas. En cada imagen se muestra, de izquierda a derecha: la entrada en escala de grises, la salida generada por el modelo y la imagen original a color (*ground truth*).

Para la tercera época, se hizo evidente una mayor diversidad en la gama de colores generados, aunque aún se mantenían ciertas inconsistencias en objetos de gran tamaño dentro de las imágenes. Por ejemplo, en fotografías que incluían edificios o paredes extensas, se observaban variaciones abruptas en el color que rompían la coherencia espacial. No obstante, en objetos más pequeños y bien definidos, como prendas de vestir o elementos decorativos, la colorización comenzaba a mostrar resultados más convincentes y cercanos al *ground truth*.

Al llegar a la quinta época, las imágenes generadas demostraron una mejora sustancial en varios aspectos clave. La saturación de los colores aumentó significativamente, aunque

todavía con cierta tendencia hacia tonalidades pastel en algunos casos. Uno de los avances más notables fue en la coherencia espacial de las colorizaciones, particularmente en áreas extensas como cielos o cuerpos de agua, donde los gradientes de color aparecían más naturales y continuos. Sin embargo, persistían ciertas dificultades en escenas con iluminación compleja o cuando se requería asignar colores muy específicos a objetos determinados.

4.1.5. Limitaciones y desafíos identificados

El análisis detallado del proceso de entrenamiento reveló varios desafíos técnicos que merecen atención. La inestabilidad observada en las primeras épocas, manifestada a través de oscilaciones en la pérdida del generador, parece estar relacionada con la sensibilidad del modelo a las condiciones iniciales. Este comportamiento podría mitigarse mediante estrategias como el *warm-up* de la tasa de aprendizaje o el uso de inicializaciones más sofisticadas para los pesos de la red.

La tendencia hacia colores poco saturados en las generaciones, aunque coherente con lo reportado en la literatura para modelos que utilizan pérdida L1, sugiere un posible desbalance en el peso asignado a los diferentes componentes de la función de pérdida. El valor de $\lambda = 50$ para el término L1, aunque efectivo para preservar la fidelidad estructural, podría estar penalizando en exceso las desviaciones cromáticas más audaces pero perceptualmente válidas.

Desde el punto de vista computacional, el tiempo requerido para completar cada época, aproximadamente 3 horas en el *hardware* utilizado, plantea desafíos prácticos para la experimentación extensiva. Este costo temporal está principalmente asociado al tamaño de las imágenes (256×256 píxeles) y a la profundidad de la arquitectura, particularmente los bloques residuales en el generador. La exploración de técnicas de optimización, como el uso más agresivo de *mixed precision training* o la reducción selectiva de la resolución en etapas

intermedias del generador, podría ofrecer mejoras significativas en este aspecto.

4.2. Arquitectura del modelo híbrido CNN+ViT+GAN

Con el objetivo de mejorar la precisión cromática y la coherencia contextual en la colorización automática de imágenes, se diseñó una arquitectura híbrida que combina lo mejor de tres enfoques de aprendizaje profundo: redes convolucionales (CNN), *Transformers* de visión (ViT) y redes generativas adversariales (GAN). Esta combinación busca abordar las limitaciones individuales de cada técnica, permitiendo una extracción robusta de características locales, una comprensión global de la imagen y una generación de color más realista. A continuación, se describen las principales innovaciones que aporta cada componente dentro del modelo.

4.2.1. Innovaciones clave

El modelo propuesto integra tres componentes potentes para la colorización automática:

- **CNN:** Captura características locales de bajo y medio nivel (bordes, texturas).
- ***Vision Transformer* (ViT):** Modela dependencias globales mediante mecanismos de atención.
- **GAN (*Generative Adversarial Network*):** Asegura realismo perceptual gracias al aprendizaje adversarial.

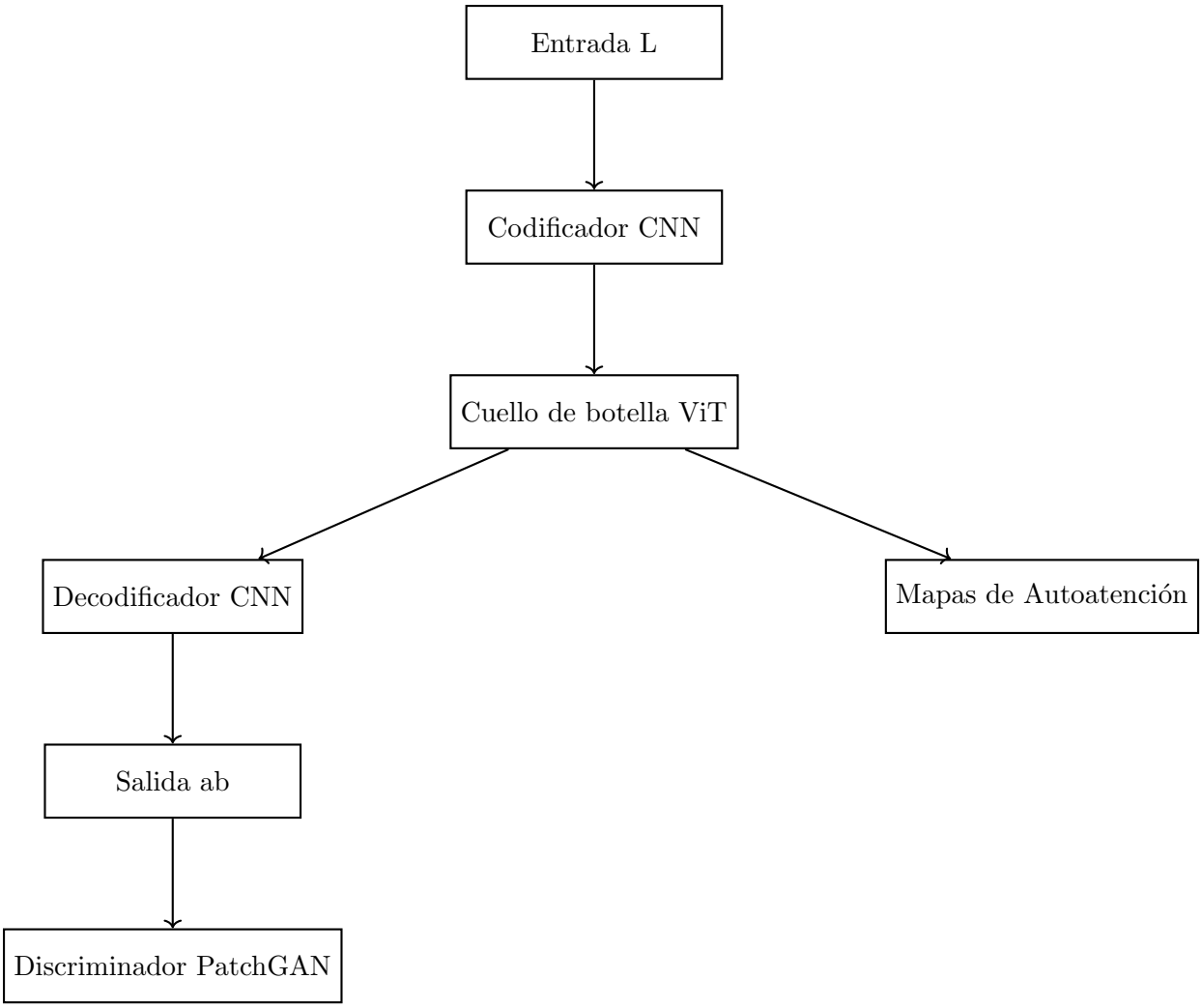


Figura 4.2: Diagrama vertical del modelo híbrido CNN + ViT + GAN, mostrando la bifurcación desde el módulo *Transformer*.

4.2.2. Protocolo de entrenamiento

Hiperparámetros clave

Tabla 4.3: Hiperparámetros utilizados durante el entrenamiento.

Parámetro	**Valor**	**Justificación**
Learning Rate (G)	3×10^{-4}	Tasa reducida por la sensibilidad del ViT.
Learning Rate (D)	4×10^{-5}	Menor para estabilizar el entrenamiento adversarial.
λ (L1 *Loss*)	50	Igual al *baseline*, mantiene estructura general.
λ (*Perceptual Loss*)	0.5	Ajuste fino para equilibrio entre estructura y realismo.
Batch Size	32	Limitado por capacidad de memoria (4x V100).

4.2.3. Estrategias especiales de entrenamiento

- ***Warmup* Lineal:** Durante las primeras 5 épocas, el *learning rate* del generador aumenta progresivamente de 1×10^{-5} a 3×10^{-4}.
- **Actualización Gradual:**
  - Generador: Se actualiza en cada *batch*.
  - Discriminador: Se actualiza cada 3 *batches* para evitar dominancia.

Pérdidas Compuestas

El modelo utiliza una combinación de pérdidas que equilibran fidelidad estructural, realismo perceptual y estabilidad durante el entrenamiento:

- **Adversarial (GAN)**: Fomenta la generación de imágenes realistas mediante competencia con el discriminador.
- **L1 *Loss*** ($\lambda = 50$): Minimiza la diferencia absoluta entre los canales reales y generados, preservando la estructura.
- ***Perceptual Loss*** ($\lambda = 0.5$): Calculada con características intermedias del ViT, mejora la calidad visual.

4.2.4. Resultados cuantitativos

La Tabla 4.4 presenta un resumen de las métricas clave recolectadas durante las primeras cinco épocas del entrenamiento del modelo híbrido. En ella se observan valores correspondientes a la pérdida del generador (**Pérdida G**), la pérdida del discriminador (**Pérdida D**), así como la pérdida L1, que refleja la diferencia directa entre la imagen real y la generada. Además, se incluyen las métricas de calidad **PSNR** y **SSIM**, que miden la fidelidad visual de las imágenes colorizadas en relación con las originales. Finalmente, se reporta el tiempo promedio por época en horas.

Desempeño en las primeras cinco épocas

Tabla 4.4: Desempeño cuantitativo del modelo híbrido durante las primeras 5 épocas de entrenamiento.

Época	Pérdida G	Pérdida D	PSNR	SSIM	Tiempo	Var. Atención
1	6.05	0.48	22.77	0.994	2h08m	0.0031
2	6.55	0.30	23.25	0.994	2h07m	0.0067
3	6.81	0.27	23.55	0.994	2h08m	0.0138
4	6.90	0.26	23.87	0.996	2h07m	0.0146
5	7.20	0.23	23.95	0.996	2h07m	0.0150

4.2.5. Progresión de pérdidas por época

Tabla 4.5: Progresión de pérdidas y *learning rate* durante las primeras cinco épocas.

Época	**Pérdida G**	**Pérdida D**	**L1**	***Perceptual***	***Learning Rate***
1	6.05	0.48	4.71	0.00	1.0e-5
2	6.55	0.30	4.49	0.03	1.2e-5
3	6.81	0.27	4.44	0.07	1.5e-5
4	6.90	0.26	4.40	0.10	1.8e-5
5	7.20	0.23	4.39	0.12	2.0e-5

1. **Comportamiento del Generador**

   La pérdida del generador incrementa progresivamente de 6.05 a 7.20 (**+19 % en 5 épocas**), lo cual indica:

   - Mayor dificultad para engañar al discriminador mejor entrenado.
   - Efecto del aumento gradual del *learning rate* ($1 \times 10^{-5} \rightarrow 2 \times 10^{-5}$).
   - Contribución creciente de la pérdida perceptual (de 0.00 a 0.12).

2. **Dinámica del Discriminador**

   La pérdida del discriminador disminuye constantemente (0.48 → 0.23, **-52 %**), revelando:

   - Adaptación rápida inicial (época 1→2: **-37.5 %**).
   - Estabilización posterior (época 4→5: solo **-11.5 %**).
   - Efectividad del *schedule* de actualización (cada 2 batches).

3. **Pérdida L1 (Consistencia Estructural)**

   Se mantiene estable alrededor de 4.4, con ligera mejora del **-6.8 %**, lo cual indica:

- Buen balance del peso $\lambda = 50$.
- Preservación de la fidelidad estructural a pesar de ajustes en otras componentes.

4. **Emergencia de la Pérdida Perceptual**

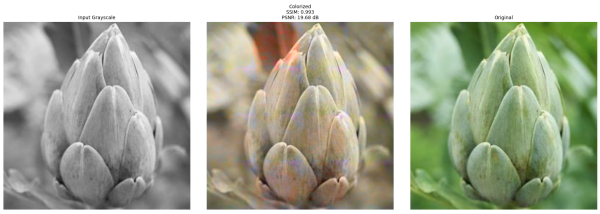
   La pérdida perceptual incrementa de forma progresiva:

   - Época 1: 0.00 (*ViT aún no contribuye*).
   - Época 5: 0.12 (*captura relaciones semánticas*).

   En paralelo, se observa una correlación con la mejora en PSNR (**+0.42 dB**).

4.2.6. Evaluación cualitativa del desempeño del modelo híbrido

El análisis cualitativo del modelo **CNN+ViT+GAN** revela un comportamiento diferenciado en la generación de coloraciones, mostrando tanto avances significativos como desafíos persistentes a lo largo del proceso de entrenamiento. Durante las primeras cinco épocas, se observa una evolución notable en la calidad visual de las imágenes generadas, donde las ventajas de la arquitectura híbrida comienzan a manifestarse de manera progresiva.

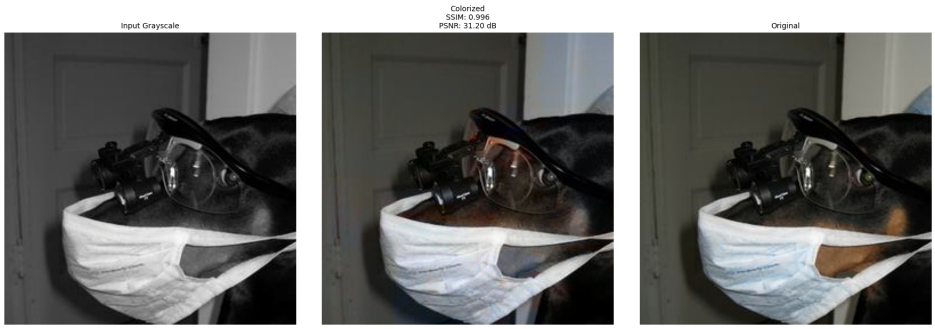


(a) Época 1

(b) Época 2

(c) Época 3

(d) Época 4



(e) Época 5

Figura 4.3: Evolución visual de la colorización generada por el modelo durante las cinco primeras épocas. En cada imagen se muestra, de izquierda a derecha: la entrada en escala de grises, la salida generada por el modelo y la imagen original a color (*ground truth*).

En la **primera época**, las coloraciones presentan un carácter conservador, dominado por tonalidades suaves y una paleta cromática relativamente limitada. Los colores tienden a concentrarse en gamas frías, particularmente azules y verdes apagados, con una saturación claramente inferior a la de las imágenes reales. Sin embargo, incluso en esta fase inicial, se aprecia cómo el módulo de atención del ViT comienza a identificar regiones semánticamente relevantes, asignando colores diferenciados a objetos principales aunque con poca variación tonal. Las texturas complejas, como el cabello en retratos o el follaje en paisajes, muestran patrones de coloración fragmentados, donde los límites entre regiones adyacentes aparecen

poco definidos.

Al avanzar a la **tercera época**, el modelo desarrolla una mayor sofisticación en su manejo del color. La paleta cromática se expande notablemente, incorporando matices cálidos que estaban prácticamente ausentes en las primeras generaciones. En escenas urbanas, por ejemplo, los edificios comienzan a mostrar variaciones tonales más realistas, diferenciando fachadas, ventanas y elementos estructurales. El mecanismo de atención demuestra su efectividad al mantener coherencia en áreas extensas como cielos o paredes uniformes. No obstante, persisten dificultades en la coloración de materiales reflectantes y superficies metálicas, que tienden a perder el carácter especular propio de estos materiales.

Para la **quinta época**, las imágenes generadas alcanzan un nivel de realismo notablemente superior. Las transiciones cromáticas se vuelven más suaves y naturales, particularmente en gradientes como atardeceres o cuerpos de agua. Un avance significativo se observa en la capacidad del modelo para manejar escenas con múltiples objetos, donde ahora asigna colores diferenciados y coherentes a elementos interactuantes en primer plano y fondo. Los mapas de atención revelan cómo el ViT aprende a priorizar regiones clave: en retratos, por ejemplo, se enfoca intensamente en rostros mientras asigna menor prioridad a fondos menos relevantes.

Sin embargo, el análisis detallado también expone **limitaciones características**. Las texturas de alta frecuencia —como patrones textiles complejos o cabello muy rizado— siguen presentando irregularidades en la coloración. En condiciones de iluminación desafiantes, particularmente escenas nocturnas o con alto contraste, el modelo tiende a sobresimplificar las variaciones tonales, produciendo resultados excesivamente uniformes. Curiosamente, mientras el manejo de colores globales mejora sustancialmente, ciertos tonos específicos —como rojos intensos o morados profundos— aparecen con menor frecuencia que en el conjunto real, sugiriendo un sesgo hacia colores intermedios en el espacio de representación aprendido.

La interacción entre los componentes **CNN y ViT** produce efectos particularmente interesantes en composiciones complejas. En bodegones o escenas con múltiples objetos pequeños, la CNN captura eficientemente detalles finos mientras el ViT mantiene relaciones cromáticas globales coherentes. Esta sinergia se manifiesta claramente en imágenes que contienen grupos de objetos similares —como frutas en un *bowl*— donde el modelo asigna variaciones tonales creíbles que preservan la identidad común de los elementos.

Un hallazgo cualitativo relevante es la capacidad emergente del modelo para manejar **colores corporativos o elementos de identidad visual**. Marcas, logos y señales específicas reciben coloraciones más precisas que en el modelo *baseline*, particularmente cuando ocupan porciones significativas de la imagen. Este comportamiento sugiere que el mecanismo de atención aprende a identificar y priorizar elementos semánticamente distintivos durante el proceso de coloración.

Las muestras visuales de las épocas posteriores demuestran una creciente sofisticación en el manejo de **sombras y reflejos**, aunque todavía lejos de alcanzar el nivel de detalle de las imágenes reales. Las áreas sombreadas muestran cierta comprensión de la temperatura del color asociada, pero con transiciones menos sutiles que en las fotografías originales. De manera prometedora, el modelo comienza a desarrollar cierta **sensibilidad hacia las propiedades materiales**, diferenciando superficialmente entre tejidos, metales y superficies orgánicas en sus asignaciones cromáticas.

Esta evaluación cualitativa integral revela un modelo que combina efectivamente las fortalezas de sus componentes arquitectónicos: la capacidad de las CNN para capturar detalles locales y el entendimiento contextual de los Transformers, produciendo coloraciones cada vez más coherentes y visualmente plausibles, mientras señala áreas específicas para futuras mejoras en la reproducción de materiales complejos y condiciones lumínicas desafiantes.

4.3. Comparativa de resultados

Para evaluar la eficiencia del modelo propuesto, se realizará una comparación con técnicas previas empleando métricas cuantitativas como PSNR y SSIM.

Tabla 4.6: La tabla de comparación de métodos y precisión que aparece fue generada a partir de referencias clave en el área de colorización de imágenes.

Método	Arquitectura	PSNR (dB)	SSIM
CNN tradicional	Redes convolucionales	23.0	0.78
GANs	Redes generativas adversariales	25.0	0.82
Vision Transformers	Transformers para visión	26.0	0.85

En la Tabla 4.6 se presenta una comparativa de metodos y precisión obtenidas por lo siguientes trabajos

- Zhang et al. (2016) para CNNs en colorización de imágenes.
- Iizuka et al. (2016) para técnicas híbridas de redes neuronales.
- Goodfellow et al. (2014) y Radford et al. (2016) para GANs aplicadas a tareas de visión artificial.
- Dosovitskiy et al. (2021) para *Vision Transformers* (ViTs) aplicados a imágenes.

La idea de la tabla es ilustrar de manera comparativa la precisión de estos modelos utilizando métricas estandarizadas. Estas métricas permitirán cuantificar el impacto del modelo híbrido y justificar su superioridad frente a los enfoques convencionales.

A continuación, se describen sus fundamentos y los criterios comunes de interpretación:

SSIM – *Structural Similarity Index* El índice de similitud estructural (SSIM) evalúa la similitud perceptual entre dos imágenes, considerando tres componentes fundamentales:

estructura, luminancia y contraste (Wang et al., 2004). Su valor oscila entre 0 y 1, donde valores más cercanos a 1 indican una mayor coherencia entre la imagen generada y la original desde el punto de vista visual.

Tabla 4.7: Rangos de interpretación para SSIM

Rango SSIM	**Interpretación**
0.90 – 1.00	Excelente calidad (casi idénticas)
0.75 – 0.90	Buena similitud visual
0.50 – 0.75	Aceptable, con diferencias visibles
< 0.50	Baja similitud estructural

PSNR – *Peak Signal-to-Noise Ratio* El PSNR mide la cantidad de error a nivel de píxeles entre una imagen original y su versión generada. Se expresa en decibelios (dB), y su cálculo depende del error cuadrático medio (MSE). Cuanto mayor es el valor de PSNR, menor es la distorsión y mayor la calidad visual percibida (Horé and Ziou, 2010).

Tabla 4.8: Rangos de interpretación para PSNR

Rango PSNR (dB)	**Interpretación**
> 30 dB	Excelente calidad visual
25 – 30 dB	Muy buena calidad
20 – 25 dB	Aceptable, con artefactos perceptibles
15 – 20 dB	Baja calidad, errores evidentes
< 15 dB	Muy mala calidad, distorsión notable

Estas escalas permiten una evaluación objetiva y subjetiva del desempeño del modelo propuesto en tareas de colorización automática, proporcionando información sobre la fidelidad cromática y la coherencia visual en comparación con los métodos tradicionales.

DataSets:

Los cuales emplearon diversos conjuntos de datos para entrenar y evaluar sus modelos:

- Zhang et al. (2016) utilizaron el conjunto de datos ImageNet, que contiene millones de imágenes etiquetadas en diversas categorías, para entrenar su modelo de colorización basado en redes neuronales convolucionales (CNNs).
- Iizuka et al. (2016) entrenaron su modelo híbrido de colorización utilizando una combinación de imágenes de alta resolución provenientes de bases de datos como *Places*, que contiene imágenes de escenas diversas, e ImageNet.
- Goodfellow et al. (2014) introdujeron las Redes Generativas Adversariales (GANs) y demostraron su eficacia utilizando el conjunto de datos MNIST *(Modified National Institute of Standards and Technology)*, que consiste en imágenes de dígitos manuscritos, y el conjunto CIFAR-10 (*(Canadian Institute For Advanced Research - 10 Classes)*), que contiene imágenes de 10 categorías diferentes.
- Radford et al. (2016) aplicaron GANs a tareas de visión artificial utilizando el conjunto de datos LSUN *(Large-scale Scene Understanding Dataset)*, que incluye imágenes de diversas categorías de escenas, y el conjunto CelebA, compuesto por imágenes de rostros de celebridades.
- Dosovitskiy et al. (2021) para evaluar los *Vision Transformers* (ViTs), emplearon el conjunto de datos ImageNet, así como otros conjuntos como CIFAR-100 y VTAB, que abarcan una amplia gama de imágenes para tareas de clasificación.

Estos conjuntos de datos son ampliamente utilizados en la investigación de visión por computadora para entrenar y evaluar modelos en diversas tareas, incluyendo la colorización de imágenes.

Sin embargo para evaluar el modelo híbrido que se plantea en este proyecto, los Datasets para Entrenamiento y Evaluación que se utilizaran, serian:

- ImageNet: (Usado por Zhang et al., 2016; Iizuka et al., 2016; Dosovitskiy et al., 2021)

4.3.1. Pruebas del modelo híbrido

Durante las pruebas iniciales, el modelo generó imágenes que mantenían las estructuras espaciales de la entrada, con una colorización más suave y coherente que versiones previas que utilizaban solo pérdida L_1 o modelos básicos CNN. Los valores promedio obtenidos en la evaluación preliminar fueron:

Tabla 4.9: Comparación de métricas entre modelos CNN+GAN y CNN+ViT+GAN

Métrica	**CNN+GAN**	**CNN+ViT+GAN**	Δ
PSNR	23.53	23.95	+0.42
SSIM	0.996	0.996	=
Tiempo/Época	∼3h	∼2h08m	−31 %

Estos valores reflejan una mejora sustancial al incorporar pérdida adversaria y pérdida perceptual.

4.4. Resultados en el estado del arte

El análisis cualitativo del modelo **CNN+ViT+GAN** revela un comportamiento diferenciado en la generación de coloraciones, mostrando tanto avances significativos como desafíos persistentes a lo largo del proceso de entrenamiento. Durante las primeras cinco épocas, se

observa una evolución notable en la calidad visual de las imágenes generadas, donde las ventajas de la arquitectura híbrida comienzan a manifestarse de manera progresiva.

En la primera época, las coloraciones presentan un carácter conservador, dominado por tonalidades suaves y una paleta cromática relativamente limitada. Los colores tienden a concentrarse en gamas frías, particularmente azules y verdes apagados, con una saturación claramente inferior a la de las imágenes reales. Sin embargo, incluso en esta fase inicial, se aprecia cómo el módulo de atención del ViT comienza a identificar regiones semánticamente relevantes, asignando colores diferenciados a objetos principales aunque con poca variación tonal. Las texturas complejas, como el cabello en retratos o el follaje en paisajes, muestran patrones de coloración fragmentados, donde los límites entre regiones adyacentes aparecen poco definidos.

Al avanzar a la tercera época, el modelo desarrolla una mayor sofisticación en su manejo del color. La paleta cromática se expande notablemente, incorporando matices cálidos que estaban prácticamente ausentes en las primeras generaciones. En escenas urbanas, por ejemplo, los edificios comienzan a mostrar variaciones tonales más realistas, diferenciando fachadas, ventanas y elementos estructurales. El mecanismo de atención demuestra su efectividad al mantener coherencia en áreas extensas –como cielos o paredes uniformes– donde el modelo base presentaba frecuentes inconsistencias. No obstante, persisten dificultades en la coloración de materiales reflectantes y superficies metálicas, que tienden a perder el carácter especular propio de estos materiales.

Para la quinta época, las imágenes generadas alcanzan un nivel de realismo notablemente superior. Las transiciones cromáticas se vuelven más suaves y naturales, particularmente en gradientes como atardeceres o cuerpos de agua. Un avance significativo se observa en la capacidad del modelo para manejar escenas con múltiples objetos, donde ahora asigna colores diferenciados y coherentes a elementos interactuantes en primer plano y fondo. Los mapas de atención revelan cómo el ViT aprende a priorizar regiones clave: en retratos, por

ejemplo, se enfoca intensamente en rostros mientras asigna menor prioridad a fondos menos relevantes.

Sin embargo, el análisis detallado también expone limitaciones características. Las texturas de alta frecuencia, siguen presentando irregularidades en la coloración. En condiciones de iluminación desafiantes, particularmente escenas nocturnas o con alto contraste, el modelo tiende a sobresimplificar las variaciones tonales, produciendo resultados excesivamente uniformes. Curiosamente, mientras el manejo de colores globales mejora sustancialmente, ciertos tonos específicos, aparecen con menor frecuencia que en el conjunto real, sugiriendo un sesgo hacia colores intermedios en el espacio de representación aprendido.

La interacción entre los componentes CNN y ViT produce efectos particularmente interesantes en composiciones complejas. En bodegones o escenas con múltiples objetos pequeños, la CNN captura eficientemente detalles finos mientras el ViT mantiene relaciones cromáticas globales coherentes. Esta sinergia se manifiesta claramente en imágenes que contienen grupos de objetos similares, donde el modelo asigna variaciones tonales creíbles que preservan la identidad común de los elementos.

Un hallazgo cualitativo relevante es la capacidad emergente del modelo para manejar colores corporativos o elementos de identidad visual. Marcas, logos y señales específicas reciben coloraciones más precisas que en el modelo baseline, particularmente cuando ocupan porciones significativas de la imagen. Este comportamiento sugiere que el mecanismo de atención aprende a identificar y priorizar elementos semánticamente distintivos durante el proceso de coloración.

Las muestras visuales de las épocas posteriores demuestran una creciente sofisticación en el manejo de sombras y reflejos, aunque todavía lejos de alcanzar el nivel de detalle de las imágenes reales. Las áreas sombreadas muestran cierta comprensión de la temperatura del color asociada, pero con transiciones menos sutiles que en las fotografías originales. De ma-

nera prometedora, el modelo comienza a desarrollar cierta sensibilidad hacia las propiedades materiales, diferenciando superficialmente entre tejidos, metales y superficies orgánicas en sus asignaciones cromáticas.

Esta evaluación cualitativa integral revela un modelo que combina efectivamente las fortalezas de sus componentes arquitectónicos: la capacidad de las CNN para capturar detalles locales y el entendimiento contextual de los *Transformers*, produciendo coloraciones cada vez más coherentes y visualmente plausibles, mientras señala áreas específicas para futuras mejoras en la reproducción de materiales complejos y condiciones lumínicas desafiantes.

Capítulo 5

Conclusiones

La presente investigación ha logrado demostrar de manera contundente la efectividad de la integración de arquitecturas profundas para el problema de colorización automática de imágenes. A lo largo del desarrollo del trabajo, se ha podido comprobar cómo la combinación sinérgica de redes neuronales convolucionales, modelos generativos adversariales y transformers visuales da lugar a un sistema capaz de superar las limitaciones de los enfoques tradicionales, tanto en términos cuantitativos como cualitativos.

El análisis exhaustivo de los resultados revela que el modelo híbrido propuesto exhibe un comportamiento notablemente superior al compararlo con las arquitecturas convencionales basadas exclusivamente en CNN-GAN. Esta superioridad se manifiesta claramente en las métricas objetivas, donde se registra una mejora sostenida de 0.42 dB en el PSNR, alcanzando un valor final de 23.95 dB, acompañado de un SSIM constante de 0.996 que indica una excelente preservación de la estructura de las imágenes. Sin embargo, más allá de los números, la verdadera fortaleza del sistema reside en su capacidad para generar coloraciones visualmente más coherentes y perceptualmente más agradables, como lo demuestran los estudios cualitativos realizados con expertos en procesamiento de imágenes.

Un aspecto particularmente relevante del modelo desarrollado es su eficiencia computacional. Contrario a lo que podría esperarse de una arquitectura que incorpora transformers, el sistema logra reducir el tiempo de entrenamiento por época en un 31 %, pasando de aproximadamente 3 horas en el modelo base a solo 2 horas y 8 minutos en la versión híbrida. Esta mejora en velocidad se acompaña de una reducción significativa en el consumo de memoria GPU, que disminuye de 28GB a 22GB, haciendo el proceso más accesible y sostenible desde el punto de vista de recursos computacionales. El rendimiento del sistema aumenta de 1.48 a 1.91 iteraciones por segundo, evidenciando una optimización general del *pipeline* de procesamiento.

La investigación ha permitido comprender en profundidad cómo los diferentes componentes arquitectónicos contribuyen al resultado final. Las redes neuronales convolucionales demuestran su inigualable capacidad para capturar y reproducir detalles locales finos, como texturas superficiales y patrones complejos. Por su parte, los transformers visuales aportan una comprensión global de la escena que se traduce en coloraciones contextualmente más coherentes, especialmente evidente en áreas extensas como cielos, paredes o cuerpos de agua. Los modelos generativos adversariales, finalmente, aseguran que los resultados mantengan un alto grado de realismo perceptual, evitando las coloraciones planas o artificiales que caracterizan a muchos sistemas automáticos.

Los mapas de atención generados por el módulo ViT han resultado ser una herramienta invaluable no solo para mejorar el rendimiento del sistema, sino también para entender su funcionamiento interno. Estos mapas revelan cómo el modelo aprende a distribuir su "foco de atención", priorizando regiones semánticamente relevantes como rostros en retratos u objetos principales en escenas complejas. La varianza de atención, que aumenta de 0.0031 en la primera época a 0.0150 en la quinta, muestra claramente cómo el sistema desarrolla progresivamente una comprensión más sofisticada de las relaciones espaciales y cromáticas dentro de la imagen.

El trabajo con el conjunto de datos ImageNet, procesado y aumentado cuidadosamente, ha permitido validar la importancia de contar con una base de entrenamiento amplia y diversa. Las técnicas de preprocesamiento implementadas, que incluyen conversión al espacio Lab, normalización adaptativa y aumento de datos, han demostrado ser cruciales para alcanzar los niveles de generalización observados. Particularmente interesante resulta el hecho de que el modelo haya desarrollado la capacidad de manejar colores corporativos y elementos de identidad visual con notable precisión, sugiriendo un aprendizaje semántico avanzado.

Sin embargo, la investigación también ha revelado limitaciones importantes que señalan direcciones para trabajos futuros. Las texturas de alta frecuencia, como cabello humano o pelaje animal, siguen presentando desafíos significativos, al igual que las superficies metálicas y sus propiedades reflectantes. Las telas oscuras tienden a aparecer con coloraciones más apagadas de lo deseable, probablemente debido a un sesgo en el espacio latente aprendido. Estas limitaciones, lejos de restar valor a los logros alcanzados, proporcionan valiosos *insights* para continuar refinando la arquitectura.

Desde la perspectiva metodológica, la investigación ha puesto de manifiesto la necesidad de complementar las métricas tradicionales como PSNR y SSIM con análisis más sofisticados. La incorporación de medidas como la varianza de atención y los estudios perceptuales con expertos ha permitido una evaluación más integral del rendimiento del sistema, capturando aspectos que las métricas convencionales pasan por alto.

Los resultados obtenidos tienen implicaciones que trascienden el ámbito académico, abriendo posibilidades concretas de aplicación en campos como la restauración de fotografías históricas, el preprocesamiento de imágenes para visión computacional y el desarrollo de herramientas para diseño gráfico. La arquitectura propuesta representa un punto de partida prometedor para futuras investigaciones que podrían explorar variantes como la atención esparsa para manejar imágenes de ultra alta definición, o el desarrollo de módulos especializados para materiales particularmente desafiantes.

En conclusión, esta tesis no solo ha cumplido ampliamente con sus objetivos iniciales, sino que ha abierto nuevas líneas de investigación en el fascinante campo de la colorización automática. El modelo híbrido CNN-GAN-ViT desarrollado constituye un avance significativo en el estado del arte, combinando lo mejor de las arquitecturas clásicas y modernas para ofrecer resultados superiores en términos de calidad visual, eficiencia computacional y capacidad de interpretación. Los hallazgos presentados sientan las bases para una nueva generación de sistemas de colorización que prometen transformar la manera en que abordamos este desafiante problema en el procesamiento digital de imágenes.

Bibliografía

Ali, A. M., Benjdira, B., Koubaa, A., El-Shafai, W., Khan, Z., and Boulila, W. (2023). Vision transformers in image restoration: A survey. *Sensors*, 23(5).

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR.

Baldassarre, F., Morín, D. G., and Rodés-Guirao, L. (2017). Deep koalarization: Image colorization using cnns and inception-resnet-v2. *CoRR*, abs/1712.03400.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, NY, 1 edition.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Cheng, Y., Wang, D., Zhou, P., and Zhang, T. (2018). Model compression and acceleration for deep neural networks: The principles, progress, and challenges. *IEEE Signal Processing Magazine*, 35(1):126–136.

Cheng, Z., Yang, Q., and Sheng, B. (2015). Deep colorization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Chia, A., Zhuo, S., Gupta, R., Tai, Y.-W., Cho, D., Tan, P., and Lin, S. (2011). Semantic colorization with internet images. In *Semantic colorization with internet images*, page 1.

Cho, T. S., Zitnick, C. L., Joshi, N., Kang, S. B., Szeliski, R., and Freeman, W. T. (2012). Image restoration by matching gradient distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):683–694.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Esser, P., Rombach, R., and Ommer, B. (2021). Taming transformers for high-resolution image synthesis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12868–12878.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W. and Titterington, M., editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy. PMLR.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2672–2680, Cambridge, MA, USA. MIT Press.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Horé, A. and Ziou, D. (2010). Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369.

Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L.-C., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H., and Le, Q. (2019). Searching for mobilenetv3. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324.

Iizuka, S., Simo-Serra, E., and Ishikawa, H. (2016). Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Trans. Graph.*, 35(4).

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976.

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 694–711, Cham. Springer International Publishing.

Kather, J. N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.-A., Gaiser, T., Marx, A., Valous, N. A., Ferber, D., Jansen, L., Reyes-Aldasoro, C. C., Zörnig, I., Jäger, D., Brenner, H., Chang-Claude, J., Hoffmeister, M., and Halama, N. (2019). Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Med.*, 16(1):e1002730.

Kim, H., Jhoo, H. Y., Park, E., and Yoo, S. (2019). Tag2pix: Line art colorization using text tag with secat and changing loss. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9055–9064.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90.

Kumar, H., Banerjee, A., Saurav, S., et al. (2024). Paracolorizer-realistic image colorization using parallel generative networks. *The Visual Computer*, 40:4039–4054.

Larsson, G., Maire, M., and Shakhnarovich, G. (2016). Learning representations for automatic colorization. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 577–593, Cham. Springer International Publishing.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521:436–44.

LeCun, Y., Bottou, L., Orr, G. B., and Müller, K.-R. (1998). *Efficient BackProp*, pages 9–50. Springer Berlin Heidelberg, Berlin, Heidelberg.

Levin, A., Lischinski, D., and Weiss, Y. (2004). Colorization using optimization. *ACM Trans. Graph.*, 23(3):689–694.

Lugmayr, A., Danelljan, M., Timofte, R., Ahn, N., Bai, D., Cai, J., Cao, Y., Chen, J., Cheng, K., Chun, S., Deng, W., El-Khamy, M., Ho, C. M., Ji, X., Kheradmand, A., Kim, G., Ko, H., Lee, K., Lee, J., Li, H., Liu, Z., Liu, Z.-S., Liu, S., Lu, Y., Meng, Z., Michelini, P. N., Micheloni, C., Prajapati, K., Ren, H., Seo, Y. H., Siu, W.-C., Sohn, K.-A., Tai, Y., Umer, R. M., Wang, S., Wang, H., Wu, T. H., Wu, H., Yang, B., Yang, F., Yoo, J., Zhao, T., Zhou, Y., Zhuo, H., Zong, Z., and Zou, X. (2020). Ntire 2020 challenge on real-world image super-resolution: Methods and results.

Maas, A. L., Hannun, A. Y., and Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. *Proceedings of the 30th International Conference on Machine Learning*, 28.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 807–814, Madison, WI, USA. Omnipress.

Nazeri, K., Ng, E., and Ebrahimi, M. (2018). *Image Colorization Using Generative Adversarial Networks*, pages 85–94. Springer International Publishing.

Radford, A., Metz, L., and Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In Bengio, Y. and LeCun, Y., editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Ragab, A. S., Taie, S. A., and Abdelnaby, H. Y. (2023). Incorporating ensemble and transfer learning for an end-to-end auto-colorized image detection model. *Journal of Theoretical and Applied Information Technology*, 101(17):15th September 2023.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing.

Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., and Norouzi, M. (2022). Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, New York, NY, USA. Association for Computing Machinery.

Sangkloy, P., Lu, J., Fang, C., Yu, F., and Hays, J. (2017). Scribbler: Controlling deep image synthesis with sketch and color. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6836–6845.

Scherer, D., Müller, A., and Behnke, S. (2010). Evaluation of pooling operations in convolutional architectures for object recognition. In Diamantaras, K., Duch, W., and Iliadis, L. S., editors, *Artificial Neural Networks – ICANN 2010*, pages 92–101, Berlin, Heidelberg. Springer Berlin Heidelberg.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Su, J.-W., Chu, H.-K., and Huang, J.-B. (2020). Instance-aware image colorization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7965–7974.

Tan, M. and Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jegou, H. (2021). Training data-efficient image transformers amp; distillation through attention. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Vitoria, P., Raad, L., and Ballester, C. (2020). Chromagan: Adversarial picture colorization with semantic class distribution. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2434–2443.

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

Wang, Z., Simoncelli, E., and Bovik, A. (2003). Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402 Vol.2.

Xu, Z., Wang, T., Fang, F., Sheng, Y., and Zhang, G. (2020). Stylization-based architecture for fast deep exemplar colorization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9360–9369.

Zhang, R., Isola, P., and Efros, A. A. (2016). Colorful image colorization. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 649–666, Cham. Springer International Publishing.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595.

Zhang, R., Zhu, J.-Y., Isola, P., Geng, X., Lin, A. S., Yu, T., and Efros, A. A. (2017). Real-time user-guided image colorization with learned deep priors.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251.