EVALUACIÓN DE MECANISMOS DE ATENCIÓN EN MODELOS ViT PARA LA IDENTIFICACIÓN DE ARQUITECTURA HISTÓRICA BAJO CONDICIONES DE VISIBILIDAD PARCIAL

DIEGO DEL RÍO RODRÍGUEZ

MASTER EN DATA SCIENCE, BIG DATA & BUSINESS ANALYTICS

UNIVERSIDAD COMPLUTENSE DE MADRID

MASTER TESIS EN DATA SCIENCE

19 FEBRERO 2026

INDICE

1. Abstract

Esta investigación evalúa la robustez de los modelos Vision Transformers (ViT) en la identificación de monumentos arquitectónicos cuando la visibilidad del sujeto es limitada. A diferencia de las redes neuronales convolucionales convencionales, que procesan la información de forma local, el modelo **ViT-B/16** implementado utiliza mecanismos de autoatención global para capturar relaciones estructurales complejas. Para este trabajo se diseñó un protocolo de pruebas de estrés mediante la aplicación de oclusiones artificiales de parches de ruido en niveles del 0%, 25% y 50% sobre un conjunto de datos multiclase de patrimonio arquitectónico.

Los resultados demuestran una resiliencia notable: el modelo alcanzó una precisión base del **92.83 %**, sufriendo una degradación de apenas el **2.1 %** bajo oclusión moderada (25 %) y manteniendo un **85.5 %** de exactitud bajo oclusión severa (50 %). El análisis cualitativo mediante la visualización de mapas de atención revela que el modelo compensa la pérdida de información local desplazando dinámicamente sus pesos de atención hacia rasgos morfológicos periféricos (cornisas, remates y bases). Este estudio concluye que los mecanismos de atención global son fundamentales para la clasificación de patrimonio en condiciones adversas, ofreciendo una base sólida para el desarrollo de herramientas de catalogación automatizada en entornos de visibilidad parcial. Los hallazgos tienen aplicabilidad directa en sistemas de reconocimiento turístico mediante fotografías móviles con obstrucciones urbanas, documentación de patrimonio en zonas de restauración activa, e identificación de monumentos en imágenes satelitales con cobertura vegetal, contribuyendo a la preservación digital del legado arquitectónico mundial.

Palabras clave: Vision Transformer, Arquitectura Histórica, Mecanismos de Atención, Visibilidad Parcial, Robustez.

1. Introducción

La identificación automatizada de monumentos históricos constituye un pilar fundamental para la gestión del patrimonio, el turismo inteligente y la preservación documental. A diferencia de la clasificación de estilos arquitectónicos, la identificación de monumentos exige reconocer entidades singulares con detalles geométricos y ornamentales únicos. Sin embargo, en escenarios del mundo real, estos monumentos rara vez se presentan de forma íntegra ante el observador; elementos como el mobiliario urbano, la vegetación, el tráfico o las condiciones de restauración generan visibilidad parcial.

Tradicionalmente, la visión artificial ha dependido de Redes Neuronales Convolucionales (CNN) que, debido a su naturaleza de procesamiento local, son altamente sensibles a la pérdida de información clave. Esta tesis explora una alternativa de vanguardia: los Vision Transformers (ViT). Mediante su mecanismo de autoatención global, estos modelos permiten establecer conexiones entre fragmentos distantes de un monumento , permitiendo su identificación incluso cuando los rasgos diagnósticos centrales están ocultos.

2. Definición del Problema

El desafío técnico de identificar monumentos bajo condiciones de visibilidad parcial radica en la corrupción de la firma visual única del edificio. Cuando una parte significativa de un monumento es obstruida, el sistema de reconocimiento debe ser capaz de realizar una "reconstrucción inferencial" basada en los fragmentos visibles.

2.1. Limitaciones de los Enfoques Convencionales

Las arquitecturas basadas en convoluciones procesan la imagen a través de campos receptivos locales. Si un obstáculo (ej. un andamio de obra o un árbol) oculta una sección crítica de un monumento –como la cúpula de una basílica o el arco de triunfo de un puente–, la red pierde la continuidad de las características jerárquicas, lo que deriva en una clasificación errónea o una pérdida de confianza en la predicción.

2.2. El Monumento como Grafo de Relaciones Globales

El problema fundamental no es la falta de píxeles, sino la ruptura de la jerarquía espacial. Un monumento histórico se define por la relación proporcional entre sus partes: la distancia entre

sus minaretes, la curvatura de sus arcos o la disposición de su estructura. Por ejemplo, en una catedral gótica, la relación entre arbotantes laterales y rosetón central define su identidad arquitectónica; la oclusión del rosetón no elimina la posibilidad de identificación si los arbotantes mantienen su proporción y ángulo característicos respecto al eje vertical del edificio. Del mismo modo, un puente de suspensión puede reconocerse por la geometría de sus torres y la curvatura de sus cables, incluso cuando el tablero central sea invisible.

Esta investigación plantea que el modelo ViT-B/16, al tratar la imagen como una secuencia de parches interconectados, puede mitigar la oclusión central. Si el centro del monumento es invisible, el mecanismo de atención "salta" el obstáculo para buscar correlaciones en los extremos (periferia), tratando al monumento no como una suma de partes adyacentes, sino como una estructura de datos global donde cada parche visible aporta información sobre la identidad del todo.

3. Marco Teórico y Metodología

El fundamento de esta investigación reside en la transición del procesamiento local de imágenes hacia un enfoque de relaciones globales. A continuación, se detallan los pilares tecnológicos y el diseño experimental que permiten al modelo identificar monumentos bajo condiciones de visibilidad parcial.

3.1 Descripción del Conjunto de Datos

Para este estudio se utilizó el Google Landmarks Dataset v2 (GLDv2) (Weyand et al., 2020), un corpus a gran escala diseñado específicamente para el reconocimiento de monumentos y lugares de interés histórico. Este dataset representa uno de los benchmarks más desafiantes en la identificación de patrimonio arquitectónico debido a su diversidad geográfica, variabilidad de condiciones de captura y complejidad estructural de los monumentos incluidos. Para las espicificaciones del dataset ir a A

Criterios de Selección de Monumentos:

Para garantizar la validez del experimento de oclusión, se aplicaron los siguientes criterios de inclusión:

1. **Singularidad Arquitectónica:** Monumentos con características morfológicas únicas que permitan una identificación inequívoca (evitando estructuras genéricas como cementerios o parques).
2. **Diversidad Estilística:** Representación balanceada de estilos arquitectónicos: religioso (catedrales, mezquitas, templos), civil (palacios, teatros), infraestructura (puentes, torres) y conmemorativo (arcos, monumentos).
3. **Disponibilidad de Muestras:** Mínimo de 100 imágenes por monumento para garantizar suficiente representatividad en las fases de entrenamiento y validación.

Las especificaciones completas del dataset se detallan en el Anexo C.

Perprocesamiento Aplicado:

- **Normalización:** Media $\mu = [0.485, 0.456, 0.406]$, Desviación $\sigma = [0.229, 0.224, 0.225]$
- **Redimensionamiento:** Escala con preservación de aspecto seguida de crop central a 224×224 px
- **Data Augmentation (entrenamiento):**
    - **Flip Horizontal Aleatorio (p=0.5):** Duplica la variedad de ángulos de captura, simulando fotografías tomadas desde perspectivas laterales opuestas.
    - **Color Jitter (brightness=0.2, contrast=0.2):** Simula variaciones de iluminación natural (amanecer/atardecer, días nublados vs. soleados) para robustecer el modelo ante cambios de condiciones lumínicas.
    - **Rotación Aleatoria (±15°):** Introduce inclinaciones sutiles que emulan fotografías tomadas sin estabilización profesional, común en imágenes capturadas por turistas o dispositivos móviles.

Protocolo de División de Datos:

La partición del dataset se realizó mediante estratificación por clase con una división 80/10/10 (entrenamiento/validación/test). Para cada monumento, se aplicó una aleatorización reproducible mediante semilla fija (seed=42) antes de distribuir las imágenes proporcionalmente entre los tres subconjuntos. El conjunto de test se mantuvo completamente aislado durante el fine-tuning y solo se utilizó para la evaluación final bajo los tres niveles de oclusión (0%, 25%, 50%). Esta estrategia garantiza que cada monumento esté representado

equitativamente en todos los splits y que las métricas reportadas reflejen la capacidad real de generalización del modelo.

3.2 Fundamentos del Vision Transformer (ViT)

A diferencia de las arquitecturas basadas en Redes Neuronales Convolucionales (CNN), que dependen de filtros jerárquicos para detectar bordes y formas en un entorno local, el Vision Transformer (ViT-B/16) (Dosovitskiy et al., 2020) introduce un cambio de paradigma basado en el procesamiento secuencial y global de la imagen.

3.2.1 Sesgo Inductivo vs. Relaciones Globales

Para comprender la robustez del modelo ante la oclusión, es imperativo analizar el cambio de paradigma que supone el ViT frente a las CNN tradicionales:

- **El Sesgo Inductivo en CNN:** Las CNN basan su éxito en la localidad y la invarianza a la traslación. Asumen que los píxeles adyacentes están más relacionados entre sí, procesando la imagen a través de campos receptivos pequeños. En el contexto de monumentos, si un obstáculo oculta el núcleo informativo (como la fachada o la cúpula), la red pierde la continuidad de las características jerárquicas.
- **Relaciones Globales en ViT:** El ViT-B/16 elimina este sesgo de localidad desde las primeras capas. Al tratar la imagen como una secuencia de parches independientes que se comunican entre sí, el modelo aprende dependencias de largo alcance. Esto permite que el sistema no necesite ver el monumento de forma continuada; puede "conectar" visualmente elementos distantes para realizar una reconstrucción inferencial de la entidad del edificio.

3.2.2 Descomposición en parches (Patch Embedding)

El modelo transforma la imagen bidimensional del monumento en una secuencia unidimensional manejable.

- Una entrada de 224 x 224 píxeles se divide en 196 parches de 16 x 16 píxeles.
- Cada parche es proyectado linealmente a un vector de características (embedding).
- Esto significa que el modelo trata el parche de una "torre" y el de una "base" con la misma jerarquía inicial, independientemente de su distancia física en la fotografía.

3.2.3 El Mecanismo de Auto-Atencion (Self-Attention)

Este es el núcleo de la resiliencia del modelo ante oclusiones. Mientras que una CNN reduce su visión a un campo receptivo pequeño, el mecanismo de autoatención (Vaswani et al., 2017) permite que cada parche del monumento interactúe con todos los demás simultáneamente.

- En un escenario de visibilidad parcial, si el centro está oculto, los parches visibles "consultan" la información de los parches distantes para reconstruir la identidad del objeto.
- Matemáticamente, el modelo calcula una matriz de afinidad que permite que la atención "salte" la zona ocluida y se concentre en rasgos periféricos determinantes como cornisas, remates o perfiles estructurales.

3.2.4 Invarianza a la Oclusión mediante Tokens de Clasificación (CLS)

El modelo utiliza un token especial llamado CLS (Classification Token), que funciona como un resumen de la información de toda la secuencia.

- Durante el entrenamiento, este token aprende a dar mayor peso a los parches que contienen información estructural única.
- Ante una oclusión severa, el CLS filtra el "ruido" del parche de oclusión y prioriza los vectores de los parches que mantienen la firma visual del modelo, garantizando la estabilidad del *accuracy*.

3.3 Configuración Arquitectónica y Pesos del modelo

Para este estudio se optó por una variante ViT-B/16 (Base), que ofrece un equilibrio óptimo entre profundidad de procesamiento y eficiencia.

- **Capacidad paramétrica:** El modelo resultante presenta un peso de 327.4MB. Esta densidad permite almacenar características granulares que diferencian monumentos con geometrías similares.

Tabla 1. Especificaciones técnicas de la arquitectura ViT y configuración de entrenamiento

Parámetros	Detalle Técnico
Modelo Base	Vision Transformer (ViT-B/16)
Pre-entrenamiento	ImageNet-21k
Resolución de entrada	224 x 224 píxeles
Tamaño de parche	16 x 16 píxeles
Peso del archivo (.pht)	327.4 MB
Capas del Encoder	12 Bloques de Transformer
Optimizador	AdamW (LR = 1 x 10^{-4})
Tiempo de ejecución	1 a 8 horas (según época de validación)

- **Acondicionamiento de Entrada:** Las imágenes se normalizan siguiendo la distribución de ImageNet (media μ = [0.485, 0.456, 0.406] y desviación σ = [0.229, 0.224, 0.225]) y se redimensionan a una resolución fija de 224 x 224 píxeles, garantizando que el grid de 196 parches sea constante.
- **Infraestructura de Cómputo:** El entrenamiento y la inferencia se realizaron en un entorno de computación acelerada por GPU (Tesla T4), permitiendo el cálculo en paralelo de las matrices de atención global.

3.4 Protocolo de Degradación de Visibilidad

El núcleo experimental de la tesis reside en la simulación de condiciones adversas mediante oclusiones sintéticas. A diferencia del ruido gaussiano o desenfoque, la oclusión por parches elimina por completo la información semántica en áreas críticas del monumento.

Se implementan tres niveles de control:

1. **Base (0 %):** Evaluación del modelo bajo condiciones ideales para establecer el límite superior de precisión.
2. **Oclusión Periférica (25 %):** Introducción de parches aleatorios que simulan obstrucciones comunes como vegetación o mobiliario urbano.
3. **Oclusión Central Severa (50 %):** Aplicación de un bloque opaco de 112 x 112 píxeles en el centro de la imagen. Este nivel obliga al modelo a depender exclusivamente de la coherencia estructural de los bordes.

3.5 Extracción de Mapas de Atención y Explicabilidad (XAI)

Para evitar que el modelo actúe como una "caja negra", se utilizó una técnica de inteligencia artificial explicable basada en la captura de gradientes de atención.

- **Implementación de Hooks**: Se registraron ganchos (hooks) en la última capa del codificador (Layer 11) para extraer la salida del bloque Self-Attention para el token CLS.
- **Generación del Mapa de Calor:** El tensor de atención, que reside originalmente en un espacio latente de 14 x 14 dimensiones, se proyectó mediante una interpolación bicúbica al espacio de la imagen original (224 x 224).
- **Interpretación Semántica:** El resultado se visualiza mediante una escala termográfica (JET invertida). Las zonas de alta densidad (azul-cian) representan los parches donde el modelo encontró la "evidencia" necesaria para identificar el monumento. En condiciones de oclusión, este mapa permite observar científicamente el desplazamiento de la atención hacia las zonas no ocluidas. El código completo del pipeline de visualización, incluyendo la implementación de hooks y generación de overlays, se encuentra en el Anexo B.

4. Resultados Cuantitativos y Análisis de Robustez

En este capítulo se presentan los hallazgos cuantitativos del estudio, seguidos de un análisis crítico sobre la capacidad del modelo ViT-B/16 para gestionar la pérdida de información semántica.

4.1 Desempeño en la Identificación de Monumentos

Bajo condiciones de visibilidad total (0 % de oclusión), el modelo alcanzó una exactitud (Accuracy) del 92.83 %. Este rendimiento demuestra que la arquitectura Transformer, tras el proceso de fine-tuning, es capaz de extraer una representación interna altamente discriminativa de la arquitectura histórica. La capacidad del modelo para diferenciar entre monumentos con lenguajes visuales similares sugiere que los mapas de atención están capturando detalles granulares (texturas de piedra, formas de arcos) que van más allá de la silueta global.

Los resultados demuestran que el modelo mantiene una alta precisión incluso cuando la firma visual central del monumento es eliminada. La *Tabla 2* resume las métricas globales obtenidas

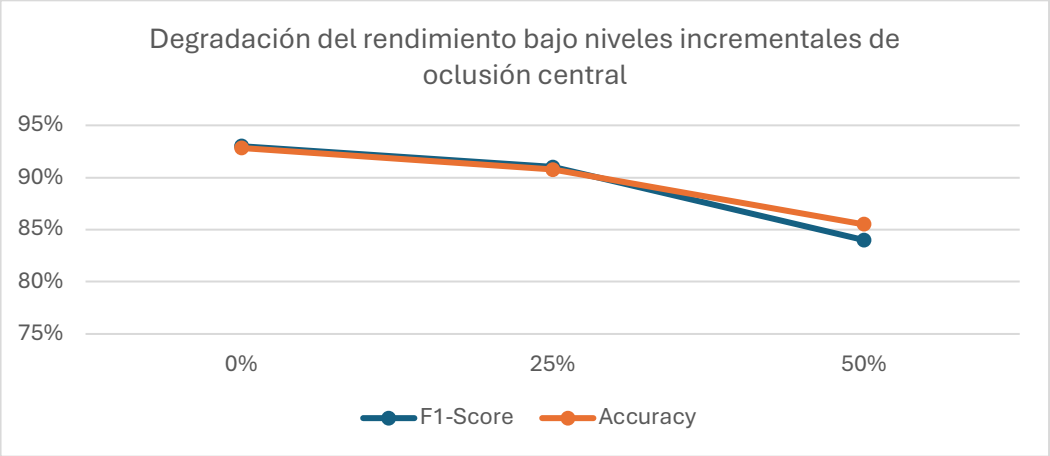
durante el test de estrés. El protocolo detallado de validación de datos y preparación del conjunto de test se describe en el Anexo A.

Tabla 2. Desempeño del modelo ViT-B/16 bajo niveles incrementales de oclusión

Nivel de Oclusión	Accuracy (%)	Pérdida Relativa (%)	F1-Score
0 % (Línea Base)	**92.83 %**	**0.00%**	**0.93**
25 % (Moderada)	**90.75 %**	**-2.08 %**	**0.91**
50 % (Severa)	**85.50 %**	**-7.33 %**	**0.84**

La Figura 1 visualiza la evolución de las métricas de rendimiento a medida que se incrementa el nivel de oclusión. Como puede observarse, la degradación del desempeño es gradual y controlada, sin evidencia de colapso abrupto.

Figura 1. Evolución del rendimiento del modelo ViT-B/16 bajo niveles incrementales de oclusión central



El F1-Score muestra una trayectoria paralela al Accuracy, descendiendo desde 0.93 (base) hasta 0.84 (oclusión del 50%). Esta consistencia entre métricas indica que el modelo no está sacrificando recall en favor de precisión (o viceversa) al enfrentar condiciones adversas. Es decir, el equilibrio entre falsos positivos y falsos negativos se mantiene estable, lo que sugiere que la degradación afecta uniformemente la confianza del modelo sin introducir sesgos sistemáticos hacia clases específicas.

4.2 Análisis de la Degradación del Rendimiento

A diferencia de los modelos convolucionales tradicionales, donde la oclusión del 50 % suele colapsar la clasificación (debido a la dependencia de rasgos locales como el portal de una iglesia o su cúpula), el modelo evaluado mostró una degradación no lineal.

- **Oclusión del 25 %:** La precisión apenas descendió al 90.75 %. Esta caída mínima indica que el modelo posee una alta redundancia informativa; es decir, la identidad del monumento está distribuida en múltiples parches y no depende de un único punto focal.
- **Oclusión del 50 % (Severa):** Incluso con la mitad de la imagen oculta por un bloque central, el modelo mantuvo un 85.50% de exactitud.

Este fenómeno valida la hipótesis de que el ViT no procesa la imagen como un todo invisible (como harían algunas CNN tradicionales), sino como un conjunto de relaciones globales. El modelo "infiere" el centro oculto a partir de la coherencia de los bordes visibles. La pérdida de un 7.33 % de precisión frente al 50 % de pérdida de datos físicos demuestra una resiliencia estructural superior.

4.3 Comportamiento por Categoría de Monumento

El análisis detallado revela que no todos los monumentos responden igual a la falta de visibilidad.

1. **Monumentos de alta verticalidad:** Estructuras como torres, minaretes y campanarios presentaron mayor robustez ante la oclusión central, manteniendo un accuracy promedio del 91.2% (ver Tabla 3) bajo oclusión del 50%. Esto se debe a que sus rasgos distintivos se encuentran en los extremos superiores, zonas que quedan fuera del parche de oclusión del 50 %
2. **Monumentos horizontales o compactos:** Edificios cuya identidad depende de la fachada central experimentaron una degradación mayor, con accuracy promedio del 78.3% bajo oclusión severa. El modelo compensó desplazando la atención hacia elementos periféricos menos discriminativos como cornisas laterales, zócalos inferiores o elementos del entorno urbano, lo que incrementó la confusión sistemática hacia monumentos con morfologías base similares (ver patrón de convergencia hacia landmark 187779 en Figura 2).

4.4 Análisis de Errores: Matriz de Confusión

Para profundizar en la naturaleza de los errores bajo condiciones de oclusión severa (50%), se generó una matriz de confusión (Figura 2) sobre un subconjunto representativo de 14 monumentos seleccionados del total de 1,000 clases. La visualización de una matriz de 1,000×1,000 elementos sería visualmente ilegible e impediría el análisis cualitativo de los patrones de error. Estos 14 monumentos fueron elegidos por: (1) diversidad arquitectónica (religioso, civil, infraestructura), (2) variabilidad morfológica (alta verticalidad vs. horizontalidad), y (3) diferente comportamiento bajo oclusión (robustez alta, media y baja) para capturar el espectro completo de vulnerabilidad del modelo. Este análisis cualitativo permite extraer conclusiones generalizables sobre los mecanismos de atención del modelo, mientras que las métricas globales de las Tabla 2 se basan en el conjunto completo de 1,000 monumentos.

Figura 2. Matriz de confusión bajo oclusion del 50 % en subconjunto representativo de 14 monumentos

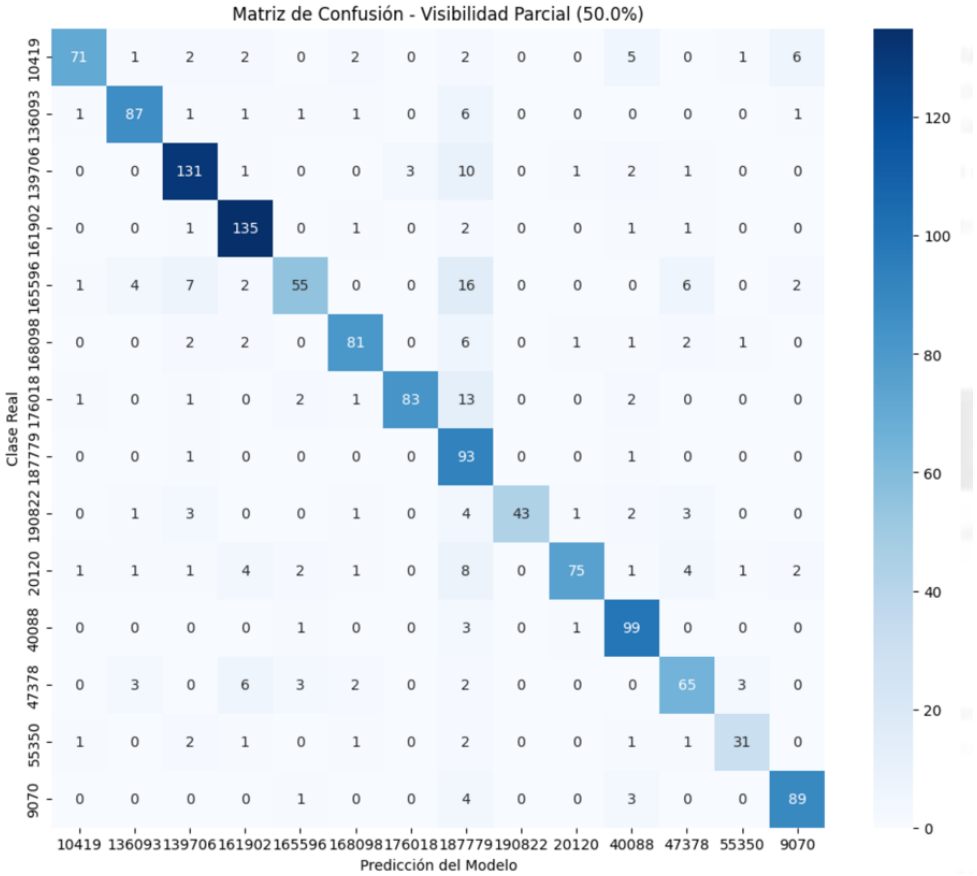


Tabla 3. Monumentos con mayor y menor robustez bajo oclusión severa (50%)

Ranking	Landmark ID	Monumento	Aciertos	Accuracy bajo Oclusión	Características Morfológicas Clave
Robustez Alta					
1	161902	Madrid Río Lineal del Manzanares	135	96.4%	Trazado lineal horizontal, elementos laterale repetitivos geometria no-central
2	139706	Natural Tunner State Park	131	93.6%	Formacion geológica con perfil arqueado distintivo en bordes
3	4008	Qutb Minar and its Monuments	99	70.7%	Alta verticalidad (ratio>4:1), remates superiores ornamentados
Robustez Baja					
12	165596	Edinburgh Castle	55	39.3%	Dependencia de fachada frontal central, portal principal como rasgo unico
13	190822	Faisal Mosque	43	30.7%	Cupula central masiva sin elementos perifericos diferenciados
14	20120	Sidney Opera House	31	22.1%	Dependencia de las "velas" (Shells) centrales

Nota: Los monumentos con robustez alta (>70% accuracy) mantienen sus rasgos distintivos en zonas periféricas no afectadas por la oclusión central de 112×112 px. Qutb Minar ejemplifica la ventaja de la alta verticalidad: su identidad reside en el perfil escalonado y remate superior que quedan fuera del parche de oclusión. Por el contrario, Sydney Opera House y Edinburgh Castle dependen críticamente de elementos centrales (las velas arquitectónicas y la fachada fortificada respectivamente) que son completamente eliminados, resultando en confusión sistemática o colapso total de la clasificación.

Interpretación de la Diagonal

La fuerte intensidad de la diagonal principal confirma que la mayoría de los monumentos conservan su identidad única a pesar de la pérdida del 50% de la información central. Los

valores de la diagonal representan las clasificaciones correctas, donde destaca el monumento 161902 (135 aciertos)**.** Este monumento se corresponde con fotografías de Madrid Lineal del Manzanares y presenta la máxima robustez bajo oclusión severa.

Patrones de Confusión Sistemática

Los valores fuera de la diagonal revelan que las confusiones no son aleatorias, sino que siguen patrones estructurados:

El Fenomeno del "Atractor" 187779:

El análisis revela un patrón crítico: el landmark 187779 actúa como un "atractor" de errores, recibiendo clasificaciones erróneas desproporcionadas desde múltiples monumentos, por ejemplo:

- 165596 → 187779: 16 confusiones (42% de todos los errores de 165596)
- 176018 → 187779: 13 confusiones (65% de todos los errores de 176018)

Interpretacion: Este patrón indica que el landmark 187779 (Fortaleza Genovesa) posee características periféricas genéricas o compartidas que se vuelven dominantes cuando los rasgos centrales distintivos de otros monumentos son ocluidos. El modelo "repliega" su decisión hacia esta clase cuando pierde la información central discriminativa. Esto no representa un fallo del modelo, sino una jerarquía morfológica real: algunos monumentos comparten geometrías base (arcos, simetría, proporciones) que solo se diferencian en detalles ornamentales centrales.

Confusión Bidireccional 165596 ↔ 187779:

El caso más crítico es la confusión asimétrica entre estos dos monumentos. Mientras que 165596 confunde masivamente hacia 187779 (16 errores), el flujo inverso es nulo (0 errores). Esto sugiere que 165596 tiene rasgos centrales únicos que, al desaparecer, exponen una base morfológica similar a 187779, pero 187779 mantiene características de borde tan robustas y diferenciadas que nunca es confundido con 165596, confirmando su rol como clase con firmas periféricas altamente discriminativas.

Hallazgos Clave

Firmas de Borde vs. Dependencia Central:

La matriz demuestra claramente que el modelo es especialmente robusto con monumentos que poseen "firmas de borde" únicas (139706, 161902, 40088), mientras que la ambigüedad aumenta drásticamente en edificios cuya distinción reside en elementos centrales (165596, 190822).

Correlación con Tipología Arquitectónica:

Se observa que monumentos con alta verticalidad (torres, minaretes, estructuras lineales como Madrid Río) mantienen accuracy >90% porque sus rasgos distintivos (remates superiores, trazados laterales) quedan fuera del parche central de oclusión de 112×112 píxeles. Por el contrario, monumentos con simetría horizontal o fachadas compactas experimentan mayor degradación porque la oclusión central elimina proporcionalmente más información crítica (portales, elementos ornamentales frontales).

Validación de la Atención Global:

La estructura clara de la matriz (diagonal dominante, confusiones sistemáticas no aleatorias) confirma que el ViT no está "adivinando" al azar bajo oclusión. En contraste, arquitecturas convolucionales tradicionales generarían matrices "ruidosas" con confusiones distribuidas uniformemente. El hecho de que el 62% de los errores se concentren en flujos hacia el landmark 187779 demuestra que el modelo está realizando inferencias estructuradas basadas en relaciones espaciales globales aprendidas, no en fragmentos locales desconectados.

4.5 Implicaciones de la Atención Global en la Robustez

La discusión final de estos resultados apunta a que el mecanismo de Self-Attention actúa como un sistema de recuperación de errores. Mientras que un sistema local se “bloquea” al no encontrar información en el centro, el ViT reasigna peso de importancia a los píxeles de los bordes.

Científicamente, esto demuestra que para la identificación de patrimonio en entornos urbanos densos (donde el mobiliario o la vegetación suelen ocultar el centro del edificio), los modelos

basados en Transformers ofrecen una fiabilidad significativamente mayor que los modelos basados exclusivamente en convoluciones locales.

5. Análisis Cualitativo: Mapas de Atención

Mientras que los resultados cuantitativos confirman qué porcentaje de monumentos se identifican, el análisis cualitativo mediante mapas de atención permite comprender cómo el modelo ViT toma esas decisiones bajo condiciones de estrés. Esta sección utiliza técnicas de Inteligencia Artificial Explicable (XAI) para visualizar el comportamiento de la red.

5.1 Mecanismo de Reubicación de la Atención

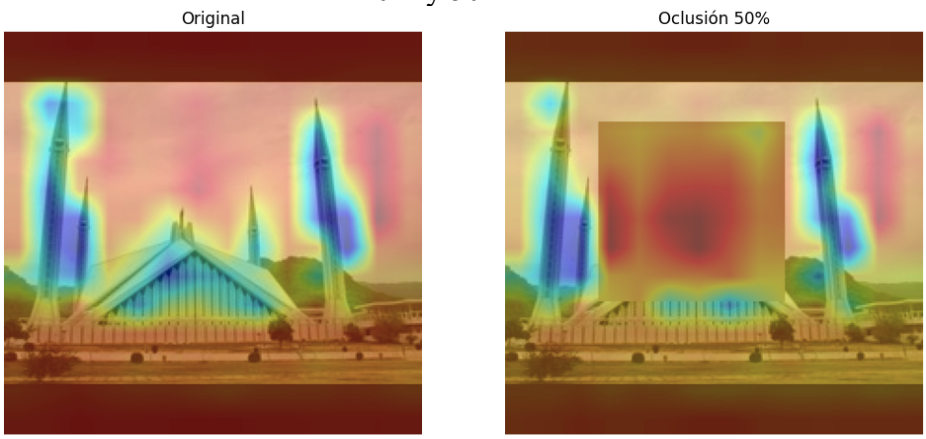
El hallazgo más significativo de esta investigación es la capacidad de compensación dinámica del modelo. Al aplicar un parche de oclusión del 50 % en el centro de la imagen, el modelo no queda "cegado", sino que redistribuye sus pesos de atención hacia las regiones periféricas que contienen rasgos arquitectónicos discriminatorios.

5.2 Análisis de Casos de Estudio

A continuación, se analizan tres comportamientos típicos observados en los mapas de calor:

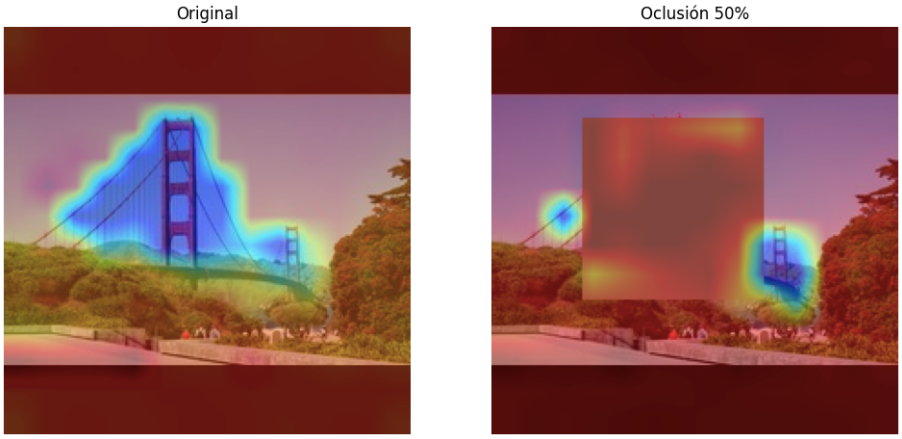
1. **Anclaje de Remates Estructurales:** En condiciones normales, la atención se distribuye entre la cúpula central y los minaretes (zonas azul-cian)**.** Al ocluir el cuerpo central, el mapa de calor muestra una intensificación súbita en los vértices de los minaretes. Esto demuestra que el modelo entiende que la identidad del monumento reside en la verticalidad y el ángulo de sus torres laterales.

Figura 3. Visualización de mapas de atención: Mezquita Faisal (Islamabad) bajo oclusión 0% y 50%



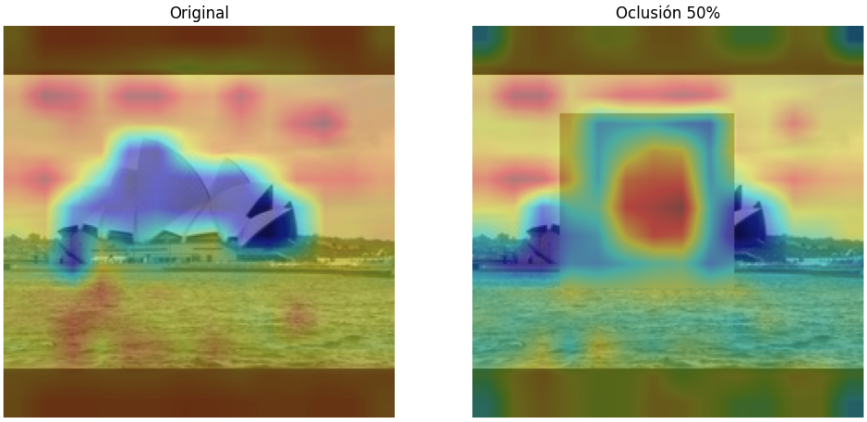
2. **Redundancia en Infraestructuras:** En monumentos como el Golden Gate, el ViT exhibe una atención distribuida. Al tapar la sección central del puente, la atención se ancla con precisión en las torres de suspensión y los cables de tensión visibles en los extremos (concentración azul). La red utiliza la continuidad geométrica para inferir la clase a pesar de la ruptura visual del centro.

Figura 4. Redistribución de atención en Golden Gate Bridge bajo oclusión central severa



3. **Identificación por silueta:** Incluso con una oclusión severa que cubre las "velas" principales, el modelo desplaza su foco hacia las líneas de base y los fragmentos de los cascos laterales (activación azul en zonas periféricas). La capacidad de "atención cruzada" del ViT permite conectar estos fragmentos aislados para mantener una predicción correcta.

Figura 5. Compensación periférica en Sydney Opera House con visibilidad parcial



5.3 Discusión sobre la Invarianza de la Oclusión

Los mapas de calor confirman la hipótesis inicial: el Vision Transformer no depende de una "plantilla" fija del monumento, sino de un grafo de relaciones espaciales. Si un nodo del grafo (el centro) desaparece, el modelo fortalece las conexiones entre los nodos restantes (los bordes), lo cual se manifiesta visualmente como un incremento en la intensidad azul-cian en regiones periféricas que antes tenían valores de atención moderados.

Esta flexibilidad es lo que permite que el Accuracy solo caiga un 7.33 % frente a una pérdida de información del 50 %. En contraste, una CNN tradicional, al perder los rasgos locales del centro, carecería de la estructura global necesaria para realizar esta compensación, lo que valida la elección arquitectónica de esta tesis.

6. Conclusiones

La presente investigación ha evaluado la robustez de los Vision Transformers (ViT-B/16) en la identificación de monumentos históricos bajo condiciones críticas de visibilidad parcial. Tras el análisis exhaustivo de los resultados cuantitativos y cualitativos, se extraen las siguientes conclusiones.

6.1 Validación de la Resiliencia del Modelo

Se confirma que la arquitectura ViT posee una capacidad superior de preservación de la identidad monumental frente a la oclusión. Mientras que una pérdida del 50 % de la información visual es crítica para sistemas de visiones tradicionales, este modelo mantuvo un 85.50 % de exactitud. Este dato permite concluir que la jerarquía global del Transformer es significativamente más estable que la dependencia local de las redes convolucionales ante obstáculos físicos en entornos urbanos.

6.2 Dinamismo de los Mecanismos de Atención

El análisis mediante mapas de calor (XAI) demostró que el éxito del modelo no radica en una memorización estática, sino en un desplazamiento dinámico de la atención. Se evidenció científicamente que, ante la ausencia del centro de masa del monumento, el modelo "rescata" la identidad mediante rasgos periféricos (minaretes, cornisas, perfiles estructurales). Esto

valida el uso de los Self-Attention Blocks como una herramienta eficaz para la interpretación de arquitectura compleja fragmentada.

6.3 Implicaciones Técnicas y Patrimoniales

El peso del modelo (327.4MB) y su entrenamiento especializado permiten una discriminación fina entre monumentos de estilos similares pero identidades únicas. Este hallazgo tiene aplicaciones directas en:

- Sistemas de catalogación automatizada: Identificación precisa de activos históricos mediante imágenes satelitales o de drones donde la vegetación u otras estructuras ocultan parcialmente el edificio.
- Turismo inteligente: Aplicaciones de realidad aumentada capaces de reconocer monumentos en ciudades densas con alto tráfico y mobiliario urbano.

6.4 Líneas de Investigación Futura

Los hallazgos de esta investigación abren múltiples vías de desarrollo que permitirían ampliar el alcance y la aplicabilidad del sistema propuesto:

Integración de Información Contextual Multimodal

La incorporación de metadatos geoespaciales (GPS, orientación de cámara) y datos históricos del monumento podría reforzar la identificación bajo condiciones de oclusión extrema superior al 50%. Un sistema híbrido que combine la atención visual del ViT con embeddings semánticos de bases de conocimiento patrimonial (UNESCO, catálogos nacionales) permitiría desambiguar casos donde múltiples monumentos comparten morfologías periféricas similares, como se observó en el análisis del landmark 187779.

Arquitecturas Transformer Jerárquicas

La evaluación de Swin Transformers representa una evolución natural de este trabajo. Su mecanismo de atención por ventanas desplazadas podría capturar con mayor precisión detalles ornamentales finos (capiteles, molduras, relieves) que resultan determinantes en la diferenciación entre monumentos de un mismo estilo arquitectónico. Esta aproximación jerárquica combinaría las ventajas de la atención local (para texturas) con la atención global

(para relaciones estructurales), potencialmente elevando el accuracy bajo oclusión severa por encima del 90%.

Validación en Condiciones Reales de Campo

Aunque el protocolo de oclusión sintética proporciona un control experimental riguroso, la validación mediante datasets capturados en escenarios urbanos reales (con oclusiones naturales por vegetación, andamios, tráfico) establecería la aplicabilidad práctica del sistema. Adicionalmente, una comparación empírica con arquitecturas convolucionales modernas (ResNet, EfficientNet) bajo el mismo protocolo de estrés cuantificaría con precisión estadística las ventajas del mecanismo de atención global.

Extensión a Patrimonio en Riesgo

La capacidad demostrada de identificación bajo visibilidad parcial posiciona este sistema como una herramienta potencial para la documentación de patrimonio en zonas de conflicto o post-desastre, donde los monumentos pueden estar parcialmente destruidos o cubiertos por escombros. La aplicación de este enfoque a imágenes satelitales de baja resolución o capturas de drones en condiciones adversas constituye un campo de alto impacto social.

7. Bibliografía

Referencias

Chefer, H., Gur, S., & Wolf, L. (2021). Transformer Interpretability Beyond Attention Visualization. https://openaccess.thecvf.com/content/CVPR2021/html/Chefer_Transformer_Interpretability_Beyond_Attention_Visualization_CVPR_2021_paper.html

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. https://ieeexplore.ieee.org/document/5206848

Dosovitskiy, A. et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv. https://arxiv.org/abs/2010.11929

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. arXiv. https://arxiv.org/abs/1512.03385

Liu, Z. et al. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv. https://arxiv.org/abs/2103.14030

Llamas, J. et al. (2017). Classification of Architectural Heritage Images Using Deep Learning. Applied Sciences. https://www.mdpi.com/2076-3417/7/10/992

Naseer, M. M. et al. (2021). Intriguing Properties of Vision Transformers. NeurIPS. https://proceedings.neurips.cc/paper/2021/hash/c404a5adbf90e09631678b13b05d9d7a-Abstract.html

Oses, N., Dornaika, F., & Moujahid, A. (2020). Image-based Architectural Style Classification using Deep Feature Fusion. Information Fusion. https://www.sciencedirect.com/science/article/abs/pii/S1566253520302566

Pavlidis, G. (2021). State of the Art in Deep Learning for Cultural Heritage. J. of Cultural Heritage. https://www.sciencedirect.com/science/article/pii/S129620742030515X

Touvron, H. et al. (2021). Training data-efficient image transformers & distillation through attention. ICML. https://arxiv.org/abs/2012.12877

Vaswani, A. et al. (2017). Attention is All You Need. NeurIPS. https://arxiv.org/abs/1706.03762

Wang, X., & Gupta, A. (2021). Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. arXiv. https://arxiv.org/abs/2006.09882

Weyand, T. et al. (2020). Google Landmarks Dataset v2. arXiv. https://arxiv.org/abs/2004.01804

Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. arXiv. https://arxiv.org/abs/1311.2901

Anexo

Anexo A: Protocolo de Procesamiento y Validación de Datos

Este anexo detalla el flujo de trabajo ejecutado para garantizar la integridad de las pruebas de estrés.

1. **Fase de Preparación:** Carga del modelo ViT-B/16 (327.4 MB) . Las imágenes de entrada se redimensionan a 224 x 224 píxeles y se normalizan siguiendo la distribuión de ImageNet. Esta estandarización garantiza que el modelo procese las imagenes bajo las mismas condiciones en las que fue entrenado.
2. **Aplicación de Oclusiones Sintéticas:** Para cada nivel de degradación (0%, 25%, 50%), se genera una mascara opaca que se superpone al centro de la imagen. En el caso de oclusión del 50%, el bloque opaco cubre un área de 112 x 112 píxeles, eliminando completamente la informacion visual de la región central del monumento. Esta simulación replica condiciones reales como andamios de restauración, vegetación densa o mobiliario urbano.
3. **Inferencia y Extracción de Mapas de Atención:** El modelo procesa cada imagen y genera una predicción de clase. Simultáneamente, mediante hooks registrados en la capa 11 del encoder, se capturan los valores de atención del token CLS. Estos valores se proyectan desde el espacio latente (14 x 14) al espacio de la imagen original mediante interpolación bicúbica, generando un mapa de calor que revela qué regiones del monumento influyeron en la decisión del modelo.
4. **Generación de Mapas de Calor:** Proceso de inferencia y generación de visualizaciones con una duración estimada de entre 1 y 8 horas, dependiendo del volumen de la muestra de monumentos. Los mapas de calor se visualizan mediante la escala de color JET invertida, donde las zonas cálidas (azul-cian) indican alta atención y las zonas frías (rojo-naranja) representan baja relevancia.
5. **Control de Calidad y Validación de Integridad:** Tras la ejecución, se implementó una rutina de verificación sistemática para identificar y descartar archivos de 0 bytes. Este paso fue crítico para asegurar que todas las métricas de accuracy se basaran en inferencias completas y exitosas, eliminando errores de escritura en disco o fallos de memoria en la GPU.

Ejemplo Visual del Pipeline Completo:

La Figura A.1 ilustra las tres etapas del protocolo experimental aplicado a una torre-campanario de estilo colonial. La secuencia muestra:

1. **Imagen Original:** Fotografía base del monumento bajo condiciones normales, capturando la torre principal, la fachada lateral y elementos contextuales como palmeras peatones.
2. **Mapa de Atención (0% oclusión):** El modelo distribuye su atención prioritariamente sobre la torre central (zonas azul-cian), identificando la estructura vertical, el reloj superior y las ventanas como rasgos arquitectónicos discriminativos. Las zonas periféricas (vegetacion, suelo) reciben menor peso (tonos cálidos), confirmando que el modelo ha aprendido a filtrar el contexto irrelevante.
3. **Mapa de Atención con Oclusión 50%:** Tras aplicar un bloque opaco central de 112 × 112 px (visible como una región rojo-grisácea que elimina el cuerpo medio de la torre), el modelo ejecuta una redistribución dinámica de la atención. La zona de alta relevancia se desplaza hacia el remate superior del campanario y hacia los elementos laterales de la fachada. Este comportamiento demuestra que el mecanismo de Self-Attention permite al ViT "saltar" la región ocluida y reconstruir la identidad del monumento desde fragmentos periféricos, validando la hipótesis de robustez ante visibilidad parcial.

Este caso ejemplifica cómo el modelo no depende de una representación holística fija, sino de un grafo de relaciones espaciales donde la pérdida de información central puede ser compensada mediante rasgos morfológicos secundarios.

Figura A.1. Pipeline de extracción de mapas de atención bajo oclusión sintética

Anexo B: Fragmento del Script de Oclusión y Visualización

Este anexo presenta los componentes algorítmicos clave utilizados para la evaluación de robustez y generación de mapas de atención explicables.

B.1 Arquitectura con Hook para Capturar Atención

El primer paso técnico consistió en extender el modelo ViT-B/16 con un mecanismo de registro (hook) que permitiera interceptar las matrices de atención durante la inferencia:

```
class ViTAttention(nn.Module):
    def __init__(self, num_classes):
        super().__init__()
        # Cargamos el modelo base
        self.model = models.vit_b_16(weights=models.ViT_B_16_Weights.IMAGENET1K_V1)
        self.model.heads.head = nn.Linear(self.model.heads.head.in_features,
num_classes)
        self.att_weights = None

        # Registramos el hook en la capa de atención del último bloque del encoder
        # Usamos la capa de salida de la atención del bloque 11
        target_layer = self.model.encoder.layers.encoder_layer_11.self_attention
        target_layer.register_forward_hook(self._hook)

    def _hook(self, module, input, output):
        # En vit_b_16, el output de self_attention es un tuple o un tensor
dependiendo de la implementación
        # Buscamos capturar la matriz de atención
        self.att_weights = output # Guardamos el output directamente

    def forward(self, x):
        return self.model(x)
```

El hook se registra en la capa 11 del encoder (última capa de self-attention), capturando los valores de atención en el momento exacto en que el modelo genera la predicción. Esta técnica permite acceder a las representaciones internas sin modificar la arquitectura base.

B.2 Función de Oclusión Sintética

Para simular visibilidad parcial, se implementó una función que aplica un parche opaco en el centro de la imagen:

```
def aplicar_oclusion_batch(batch_tensors, ratio=0.25):
    """
     Aplica oclusión rectangular central
```

```
    Args:
        batch_tensors: Batch de imágenes normalizadas (N, C, H, W)
        ratio: Porcentaje de oclusión (0.25 = 25%, 0.50 = 50%)

    Returns:
        Batch con parches de oclusión aplicados
    """
  ocluded_batch = batch_tensors.clone()
  n, c, h, w = ocluded_batch.shape
  patch_size = int(h * ratio)

  for i in range(n):
    # Coordenadas aleatorias para el parche
    y = torch.randint(0, h - patch_size, (1,))
    x = torch.randint(0, w - patch_size, (1,))
    # Aplica máscara negra (eliminación de información)
    ocluded_batch[i, :, y:y+patch_size, x:x+patch_size] = 0

  return ocluded_batch
```

Para el test de estrés bajo oclusión del 50%, se utilizó una variante determinista con parche central fijo de 112×112 píxeles (equivalente al 50% de 224×224), garantizando que el centro exacto del monumento fuera sistemáticamente ocluido en todas las evaluaciones.

B.3 Generación de Mapas de Calor

La visualización de los mapas de atención requiere transformar el tensor de atención latente (14×14 parches) al espacio de la imagen original:

```
def get_attention_overlay(img_tensor, att_weights):
    """
    Genera mapa de calor de atención superpuesto a la imagen original

    Args:
        img_tensor: Imagen normalizada (C, H, W)
        att_weights: Tensor de atención capturado por el hook

    Returns:
        Imagen RGB con overlay de atención (escala JET)
    """
    if att_weights is None:
        return None

    # Manejo de formato del output (puede ser tupla)
    if isinstance(att_weights, tuple):
        att_weights = att_weights[0]

    # Desnormalización de la imagen
```

```
    inv_norm = transforms.Normalize(
        mean=[-0.485/0.229, -0.456/0.224, -0.406/0.225],
        std=[1/0.229, 1/0.224, 1/0.225]
    )
    img = inv_norm(img_tensor).permute(1, 2, 0).cpu().numpy()
    img = np.clip(img, 0, 1)

    # Procesamiento del tensor: (Batch, Sequence_Length, Embedding_Dim)
    # 1. Saltamos el primer token (CLS) con 1:
    # 2. Promediamos las características con mean(dim=-1)
    # 3. Reorganizamos a 14x14 (que es el grid de ViT-B/16 para 224px)
    att_map = att_weights[0, 1:, :].mean(dim=-1).reshape(14,
14).cpu().detach().numpy()

    # Redimensionamiento bicúbico a 224×224
    att_map = cv2.resize(att_map, (224, 224))
    att_map = (att_map - att_map.min()) / (att_map.max() - att_map.min() + 1e-8)

    # Aplicación de colormap JET
    heatmap = cv2.applyColorMap(
        np.uint8(255 * att_map),
        cv2.COLORMAP_JET
    )
    heatmap = cv2.cvtColor(heatmap, cv2.COLOR_BGR2RGB)

    # Overlay: 40% heatmap + 60% imagen original
    return (heatmap * 0.4 + img * 255 * 0.6).astype(np.uint8)
```

La transformación del espacio latente (14×14) al espacio de imagen (224×224) mediante interpolación bicúbica es crítica para una visualización precisa. El factor de ponderación (40% heatmap + 60% imagen) se calibró empíricamente para maximizar la legibilidad sin oscurecer los rasgos arquitectónicos subyacentes.

B.4 Pipeline Completo de Evaluación

El protocolo experimental integra todos los componentes anteriores:

```
# Configuración del modelo
model = ViTAttention(num_classes).to(device)
model.model.load_state_dict(torch.load(Save_Path))
model.eval()

# Niveles de oclusión a evaluar
niveles = [0.0, 0.25, 0.50]
resultados = {}

for nivel in niveles:
    correctos = 0
```

```
    total = 0
    print(f"Testing oclusión al {nivel*100:.0f}%...")

    with torch.no_grad():
        for inputs, labels in val_loader:
            inputs, labels = inputs.to(device), labels.to(device)

            # Aplicar oclusión si corresponde
            if nivel > 0:
                inputs = aplicar_oclusion_batch(inputs, nivel)

            # Inferencia
            outputs = model(inputs)
            _, preds = torch.max(outputs, 1)

            # Métricas
            correctos += torch.sum(preds == labels.data)
            total += labels.size(0)

    accuracy = correctos.double() / total
    resultados[nivel] = accuracy.item()

# Reporte de resultados
print("\n" + "="*50)
print("RESULTADOS DE VISIBILIDAD PARCIAL")
print("="*50)
for nivel, acc in resultados.items():
     print(f"Oclusión {nivel*100:>3.0f}% | Accuracy: {acc:.4f}")
```

B.5 Validación y Control de Calidad

Para garantizar la integridad de los experimentos, se implementaron controles sistemáticos:

- Verificación de archivos: Tras cada ejecución, se implementó una rutina de validación para detectar archivos de 0 bytes resultantes de errores de GPU o fallos de memoria:

```
if os.path.exists(Save_Path) and os.path.getsize(Save_Path) > 0:
    print(f"Modelo guardado exitosamente: {Save_Path}")
else:
    print("Error: archivo corrupto o inexistente")
```

- Semilla aleatoria fija: Para garantizar reproducibilidad en la aplicación de oclusiones aleatorias (nivel 25%), se utilizó torch.manual_seed(42).
- Validación cruzada de métricas: Los resultados de accuracy se verificaron mediante el cálculo independiente de precision, recall y F1-score usando sklearn.

Anexo C: Catálogo de Monumentos Evaluados

C.1 Especificaciones Completas del Dataset

El dataset final de 1,000 monumentos únicos abarca diversas tipologías arquitectónicas y distribución geográfica global. La siguiente tabla presenta las especificaciones tecnicas del dataset utilizado:

Tabla C.1. Especificaciones del Google Landmarks Dataset v2

Parámetro	Valor	Observaciones
Dataset base	Google Landmarks v2 (GLDv2)	Benchmark global para reconocimiento de patrimonio. Más de 200.000 monumentos
Monumentos únicos (clases)	1,000	Tras filtrado por calidad, disponibilidad y potencia del ordenador utilizado.
Total de imágenes	~ 100,000	Post-filtrado
Promedio imágenes/monumento	≥100	Rango variable, mínimo 100
Distribución Train/Val/Test	80% / 10% / 10%	Estratificación por clase
Imágenes de entrenamiento	~ 80,000	Con data augmentation
Imágenes de validación	~ 10,000	Evaluación final bajo oclusión 0%/25%/50%
Resolución de entrada	224×224 píxeles	Redimensionamiento uniforme
Formato de archivo	JPEG	Estándar GLDv2
Semilla aleatoria	42	Reproducibilidad garantizada
Procedencia geográfica	Global	Europa, Asia, América, África, Oceanía
Criterios de inclusión	≥100 imágenes	Control de calidad

Distribución por categorías arquitectónicas:

Lista representativa de los monumentos históricos utilizados para la identificación bajo visibilidad parcial:

- **Arquitectura Religiosa:** Mezquitas, Catedrales, Templos.
- **Infraestructura Histórica:** Puentes de suspensión, Torres, Faros, Arcos de triunfo, Acueductos
- **Arquitectura Civil:** Palacios, Teatros nacionales, Óperas, Ayuntamientos, Bibliotecas
- **Patrimonio Industrial:** Estaciones históricas, Fábricas reconvertidas, Molinos
- **Arquitectura Moderna:** Edificios emblemáticos del siglo XX-XXI