Aprendizaje de estilo visual: métricas de similitud y conjuntos de datos de referencia

David Erroz[1], Roberto González[1], Rubén Pascual[1], Mikel Sesma-Sara[1], and Mikel Galar[1]

Institute of Smart Cities (ISC), Dept. de Estadística, Informática y Matemáticas, Universidad Pública de Navarra (UPNA), Campus de Arrosadia, 31006, Pamplona, España
{david.erroz, roberto.gonzalezmiranda, ruben.pascual, mikel.sesma, mikel.galar}@unavarra.es

Abstract. La similitud de estilo visual busca comparar imágenes según su apariencia, minimizando la sensibilidad al contenido semántico, una capacidad cada vez más relevante para recuperación por estilo, atribución y auditoría de generadores modernos de texto a imagen. Este artículo revisa el panorama actual de la similitud de estilo visual, examinando conjuntos de datos utilizados habitualmente y organizando las métricas existentes, desde aproximaciones perceptuales genéricas hasta representaciones aprendidas específicas de estilo. Para complementar la revisión, realizamos una comparación empírica unificada de métodos representativos sobre una colección diversa de conjuntos de evaluación. Mediante rendimiento en recuperación y una métrica de error de estructura de similitud, analizamos fuga de contenido y robustez entre dominios. Los resultados muestran que los métodos dedicados al estilo superan en promedio a los métodos de referencia genéricos; sin embargo, el rendimiento depende fuertemente del conjunto de datos y se degrada en colecciones reales exigentes, indicando que la similitud de estilo fiable y alineada con el juicio humano sigue siendo un problema abierto.

Keywords: Similitud de estilo visual · Desentrelazamiento estilo–contenido · Conjuntos de datos controlados.

1 Introducción

Medir la similitud de estilo visual entre imágenes es cada vez más importante en visión por computador y multimedia. A diferencia del contenido semántico, el estilo se refiere a la apariencia visual distintiva de una imagen—patrones de pincelada, paletas de color, textura, convenciones de renderizado y composición—que puede hacer que dos imágenes resulten similares pese a representar sujetos distintos. En imágenes artísticas, estas características suelen asociarse con artistas, movimientos o géneros, aunque la noción de estilo visual se aplica también a otros dominios. Las medidas fiables de similitud de estilo son valiosas para recuperación basada en estilo [31], asistencia creativa [38] y control de calidad en procesos de diseño donde debe mantenerse una identidad visual consistente [35].

Esta necesidad se ve amplificada por los generadores modernos de texto a imagen [28], capaces de imitar artistas o géneros visuales a gran escala. Sin embargo, el análisis computacional del estilo sigue siendo desafiante: el estilo es multifactorial y subjetivo, y la frontera entre contenido y estilo suele ser ambigua [31,35]. Recientemente se han propuesto nuevos conjuntos para enfatizar variación estilística y evaluación controlada [35,23], junto con métodos que aprenden representaciones de estilo para recuperación y atribución en imágenes reales y generadas [31,35,41,17]. Esto hace oportuno consolidar qué se ha propuesto, cómo se evalúa y qué permanece abierto.

Nos centramos en la similitud de estilo de grano fino, es decir, proximidad estilística más allá de etiquetas categóricas amplias, que a menudo son demasiado gruesas para reflejar variaciones sutiles relevantes en recuperación, atribución y evaluación de modelos generativos. Una métrica efectiva debería enfatizar pistas estilísticas, reducir la sensibilidad al contenido semántico y la composición, y seguir siendo significativa entre dominios visuales heterogéneos.

Este artículo revisa conjuntos de datos y métricas de similitud de estilo visual de grano fino (Secciones 2–3) y compara métodos representativos bajo un protocolo común para analizar fuga de contenido, sensibilidad estilística y robustez entre dominios (Sección 5). Nuestras contribuciones son: (i) una revisión de conjuntos de datos para análisis de estilo visual; (ii) una organización de métricas de similitud de estilo; y (iii) una comparación unificada de métodos representativos.

2 Conjuntos de datos

El progreso en comprensión del estilo visual y recuperación basada en estilo depende de cómo se anota el "estilo" y de su entrelazamiento con el contenido. A diferencia de conjuntos centrados en objetos, el estilo visual rara vez es una etiqueta objetiva: puede ser categórico (movimiento artístico), basado en identidad (artista), débilmente supervisado (etiquetas, descripciones textuales o coocurrencia en carteras de obras), o sintético y controlado. Resumimos los conjuntos más usados según su régimen de supervisión.

2.1 Colecciones artísticas curadas

Las colecciones artísticas curadas proporcionan metadatos relativamente fiables (artista, movimiento/estilo, periodo, materiales), aunque sus anotaciones suelen ser gruesas y correlacionarse con temas recurrentes. *WikiArt* es el estándar de facto, con etiquetas de artista, estilo/movimiento y género; el subconjunto habitual en [32] incluye más de 80.000 pinturas de 1.119 artistas y 27 estilos. *WikiStyle+* [19] extiende *WikiArt* con descripciones textuales de estilo y contenido, permitiendo entrenar y evaluar representaciones condicionadas por texto y analizar desentrelazamiento contenido–estilo. *ArtBench-10* [18] ofrece un conjunto limpio y balanceado de 10 estilos. *Ukiyo-e* [15] proporciona impresiones xilográficas a gran escala con metadatos, y *SemArt* [7] empareja pinturas con comentarios y atributos textuales.

Otros conjuntos agregados combinan obras de varias colecciones y metadatos de catálogo; destacan *Art500k* [21] y *OmniArt* [37]. Conjuntos museísticos como *Rijksmuseum Challenge* [22] permiten análisis por artista o periodo, aunque no están diseñados alrededor de una taxonomía de estilo de grano fino. En cambio, *MET* [43] y *OpenMIC* [16] están orientados a instancias individuales y son menos directos para evaluar similitud de *estilo*, aunque útiles para patrimonio cultural.

También existen conjuntos especializados y comunitarios. *Artistic-Faces* [42] se centra en retratos con puntos de referencia faciales, mientras que *StyleBabel* [30] ofrece etiquetas libres de estilo y descripciones anotadas por expertos y colaboradores no expertos, permitiendo recuperación de estilo imagen–imagen e imagen–texto. Conjuntos de *Kaggle/HuggingFace* suelen derivar de fuentes tipo WikiArt (*Painting-91* [14], *Painter by Numbers* [13], *MultitaskPainting100k* [3], *Artistic Styles* [1], *WikiArt Movements* [27], *Painting-style-classification* [2], *Impressionist_Classifier_Data* [5], *Artistic Painting* [25] y *Best Artworks of All Time* [11]), heredando etiquetas gruesas y entrelazamiento contenido–estilo. Finalmente, *APDDv2* [12] y *LAPIS* [20] se centran en calidad estética y valoraciones subjetivas, útiles para estudiar confusión entre pistas estéticas y estilísticas.

2.2 Conjuntos recopilados de la web y débilmente supervisados

Estos conjuntos aumentan la escala y diversidad, aunque sus etiquetas suelen ser ruidosas, ambiguas y multietiqueta. *ContraStyles* [35] se construye filtrando *LAION-Aesthetics 6+* [33] con miles de etiquetas de estilo, produciendo un gran conjunto para aprendizaje contrastivo de representaciones. *BAM* y *BAM-FG* [40,31] agregan obras generadas por usuarios con pistas de estilo débiles y supervisión basada en agrupaciones. *NoisyArt* [4] aborda reconocimiento de arte con supervisión procedente de la web.

Más allá de arte, *DomainNet* [24] evalúa comportamiento entre dominios compartiendo categorías de contenido (pintura, boceto, clipart), aunque sus etiquetas actúan como aproximaciones gruesas de estilo. *Stylebreeder* [45] recopila millones de imágenes e instrucciones textuales generadas por usuarios sin etiquetas explícitas de estilo.

2.3 Conjuntos de evaluación controlados y sintéticos

Los conjuntos controlados desacoplan explícitamente estilo y contenido, por lo que son especialmente informativos para diagnosticar fuga de contenido. *fruitSALAD* [23] varía independientemente conceptos de fruta y estilos en un diseño 10×10. *OmniConsistency* [36] proporciona estilizaciones emparejadas en 22 estilos. *ArtSplit* [17] sintetiza combinaciones estilo–semántica mediante instrucciones textuales separadas y *Stable Diffusion*.

3 Métricas de estilo

Evaluar similitud de estilo visual es desafiante porque el estilo es multifactorial y subjetivo. Idealmente, una métrica debería asignar alta similitud a imágenes con características estilísticas comunes independientemente del contenido, y baja similitud cuando los estilos divergen. Los enfoques existentes se agrupan en métricas perceptuales genéricas y representaciones específicas de estilo.

3.1 Métricas genéricas de similitud

Los métodos de referencia comunes reutilizan medidas perceptuales no diseñadas para aislar estilo. Métodos *IQA* de referencia completa como *SSIM* [39], *LPIPS* [44] y *DISTS* [6] comparan estructura, características profundas o estadísticas de textura. Métricas distribucionales como *FID* [10] comparan distribuciones de características, pero no producen una distancia consulta–referencia directa. Representaciones auto-supervisadas como *DINOv3* [34] se usan para comparación imagen–imagen mediante similitud coseno. Asimismo, *CLIPScore* [9], propuesto para alineamiento imagen–texto con CLIP [26], se reutiliza comparando *embeddings* visuales (*CLIP-I*) [29]. En conjunto, estas métricas tienden a confundir estilo y contenido, motivando métodos específicos de estilo.

3.2 Métricas específicas de estilo

Las métricas específicas mapean una imagen a una representación de estilo destinada a capturar apariencia (paleta, textura, pinceladas o renderizado) con menor sensibilidad al contenido. Esta idea fue popularizada por *Neural Style Transfer (NST)* [8], donde las matrices de Gram de características CNN capturan estadísticas de textura/color. Trabajos recientes reutilizan modelos generativos: *IntroStyle* [17] compara características intermedias de modelos de difusión, y *WikiStyle+* [19] aprende códigos contenido/estilo alineando imágenes con descripciones textuales. Sin embargo, estas representaciones no se optimizan explícitamente para ordenación discriminativa, motivando métodos supervisados.

ALADIN. *ALADIN* [31] aprende un *embedding* discriminativo agregando estadísticas AdaIN de múltiples capas en un *encoder–decoder*, entrenado con supervisión débil de BAM-FG. *ALADIN-ViT* [30] sustituye el *encoder* CNN por un ViT manteniendo el protocolo basado en BAM-FG.

CSD. *CSD* [35] aprende un *embedding* ViT con pérdida contrastiva multietiqueta sobre *ContraStyles* y un término auto-supervisado preservador de estilo que evita aumentos fotométricos y usa transformaciones espaciales para fomentar invariancia a la composición reteniendo textura y paleta.

GOYA. *GOYA* [41] aprende espacios contenido/estilo usando imágenes sintéticas de *Stable Diffusion* generadas con *prompts* que especifican ambos atributos. Codifica imágenes con CLIP congelado y entrena dos MLPs ligeros con objetivos contrastivos: positivos de contenido por similitud CLIP-texto y positivos de estilo por mismo estilo de *prompt*, además de una pérdida auxiliar de clasificación de estilo.

4 Configuración experimental

Evaluamos métodos representativos en un protocolo unificado sobre conjuntos con distintos regímenes y dominios para caracterizar fuga de contenido, sensibilidad estilística y robustez entre dominios.

4.1 Métodos de similitud de estilo

Evaluamos métodos representativos.[1] Comparamos *NST Gram*, *IntroStyle*, *ALADIN (ALADIN-ViT)*, *CSD (ViT-L)* y *GOYA*, con *CLIP-I (CLIP ViT-L/14)* y *DINOv3 (ViT-L/16)* como métodos de referencia genéricos. Para métodos basados en *embeddings* usamos similitud coseno $s_{i,j} \in [-1, 1]$. *NST Gram* e *IntroStyle* se formulan como distancias, usadas directamente cuando basta un *ranking* específico del método. Cuando se requiere una escala común, definimos $s_{i,j} \in [-1, 1]$ mapeando sus distancias con una escala fija $d_{\max}$, estimada como percentil 99 de pares aleatorios en una partición de calibración independiente y mantenida constante en todos los conjuntos para evitar utilizar información del conjunto de evaluación durante la normalización. El preprocesado sigue las implementaciones públicas.

4.2 Conjuntos de datos

Evaluamos una colección amplia de conjuntos de datos con distintos dominios y granularidades (Tabla 1).[2] Excluimos conjuntos sin etiqueta de estilo utilizable o no disponibles. Si existen etiquetas de estilo/movimiento, las usamos como objetivo; en caso contrario, usamos artista como aproximación al estilo [35]. Usamos particiones oficiales de evaluación cuando existen; si no, el conjunto completo. Cuando hay etiquetas de concepto, retenemos una única imagen por combinación estilo–concepto para evitar que contenidos repetidos favorezcan artificialmente la recuperación por estilo. Para mitigar la variabilidad introducida por esta selección, repetimos el proceso en 20 muestreos independientes y reportamos el promedio.

4.3 Métricas de evaluación

Reportamos dos métricas complementarias.

mAP@5: cada conjunto de datos se divide en 80% base de datos y 20% consultas, estratificando para que cada etiqueta de consulta aparezca en la base.

[1] Para evitar confusión con las métricas de evaluación, nos referimos a los enfoques comparados como "métodos".

[2] Abreviaturas: WA-S (WikiArt estilo), WA-A (WikiArt artista), AB-10 (ArtBench-10), U-e (Ukiyo-e), SA (SemArt), Rijks (Rijksmuseum), m-MET (mini-MET), AF (Artistic-Faces), PSC (painting-style-classification), ICD (Impressionist_Classifier_Data), AP (Artistic Painting), AS (Artistic Styles), BA (Best Artworks), DN (DomainNet), f-SALAD (fruit-SALAD), OC (OmniConsistency).

	Curados																Déb. sup.	Controlados	
Conjunto de datos	**WA-S**	**WA-A**	**AB-10**	**U-e**	**SA**	**Rijks**	**m-MET**	**Art500k**	**APDDv2**	**LAPIS**	**AF**	**PSC**	**ICD**	**AP**	**AS**	**BA**	**DN**	**f-SALAD**	**OC**
#Imgs	80.096	80.096	10.000	17.790	1.031	18.896	38.307	43.453	1.002	11.763	160	629	990	4.102	511	8.446	2.070	100	2.605
Etiqueta	S	A	S	A	A	A	ID	A	S	S	A	S	A	A	S	A	G	S	S
#Estilos	27	1.119	10	149	628	3.028	33.501	4.992	8	26	16	26	10	20	9	50	6	10	22
Concepto	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Sí	Sí	No

Table 1: Conjuntos de datos usados tras preprocesado. #Imgs/#Estilos: número final de imágenes/clases. Etiqueta: S=estilo, A=artista, G=dominio/género, ID=obra a nivel de instancia. Concepto: disponibilidad de etiqueta de contenido.

	Curados																Déb. sup.	Controlados		
	WA-S	**WA-A**	**AB-10**	**U-e**	**SA**	**Rijks**	**m-MET**	**Art500k**	**APDDv2**	**LAPIS**	**AF**	**PSC**	**ICD**	**AP**	**AS**	**BA**	**DN***	**f-SALAD***	**OC**	**Media**
NST Gram	,33/,41	,28/,39	,37/,42	,41/,39	,06/,41	,31/,44	,17/,40	,16/,39	,42/,41	,41/,42	,53/,40	,17/,41	,46/,42	,59/,40	,41/,33	,51/,40	,66/,30	,55/,35	,83/,24	,40/,39
ALADIN	,48/,24	,33/**,21**	,46/,24	,39/,22	,05/,21	,32/,25	,16/,07	,14/,22	,51/,25	,48/,24	,67/,15	,26/,23	,46/,24	,68/**,21**	,60/,29	,59/**,21**	**,74**/,30	**,98**/**,10**	,85/,14	,48/,21
CSD	,67/,32	**,52**/,22	,63/,29	**,46**/,22	**,15**/**,14**	**,37**/,20	,22/,06	,26/**,18**	,59/,28	,62/,27	,78/**,13**	**,41**/,32	**,69**/**,23**	**,85**/,22	**,67**/,30	**,80**/,21	,71/,28	,92/,10	**,93**/**,14**	**,59**/,22
GOYA	,50/**,22**	,28/,21	,51/,23	,32/,24	,07/,20	,21/,23	,14/,13	,13/,22	,50/**,24**	,50/,22	,68/,15	,27/**,22**	,47/,24	,65/,22	,50/,23	,58/,21	,62/,23	,71/,17	,68/,20	,44/**,21**
CLIP-I	**,68**/,23	,51/,21	**,65**/**,23**	,43/,30	,14/,21	,28/,23	,21/,17	,25/,22	**,60**/,24	**,64**/**,22**	**,79**/,22	,40/,23	,68/,25	,82/,22	,63/**,22**	,78/,22	,59/**,23**	,55/,25	,76/,20	,55/,23
DINO	,60/,29	,42/,25	,56/,28	,44/**,20**	,12/,16	,27/**,20**	**,24**/**,04**	**,27**/,24	,50/,33	,58/,25	,75/,14	,36/,28	,52/,27	,70/,25	,64/,32	,65/,24	,42/,38	,13/,27	,47/,28	,45/,25
IntroStyle	,53/,49	,36/,47	,50/,48	,41/,33	,09/,29	,30/,40	,15/,12	,19/,45	,47/,48	,52/,49	,48/,38	,27/,47	,52/,46	,68/,46	,53/,48	,61/,46	,65/,45	,97/,28	,87/,38	,48/,41

Table 2: Comparación cuantitativa. Cada celda reporta mAP@5 / E_{avg}; la negrita indica el mejor valor por conjunto de datos y métrica. Para compactar la tabla, se omite el cero inicial en los valores decimales. *DN y f-SALAD: promedio de 20 ejecuciones tras seleccionar una única imagen por par estilo–concepto.

Para cada consulta recuperamos los top-5 vecinos y calculamos mAP@5 considerando relevantes las imágenes con la misma etiqueta de estilo.

Error de similitud de estilo. Comparamos la matriz de similitud predicha $s_{i,j}$ con una estructura ideal $s^*_{i,j} = 1$ si $y_i = y_j$ y 0 en caso contrario. Las similitudes negativas se truncan a 0 y se calcula MSE intra- e inter-estilo:

$$E_{\text{intra}} = \frac{1}{|\{(i,j) \mid y_i = y_j\}|} \sum_{(i,j)|y_i=y_j} (s_{i,j} - 1)^2,$$

$$E_{\text{inter}} = \frac{1}{|\{(i,j) \mid y_i \neq y_j\}|} \sum_{(i,j)|y_i\neq y_j} (s_{i,j})^2.$$

Reportamos $E_{\text{avg}} = (E_{\text{intra}} + E_{\text{inter}})/2$, donde valores menores indican mayor acuerdo con la separación ideal.

5 Resultados

Presentamos resultados cuantitativos (Tabla 2) y matrices de similitud cualitativas (Figura 1), que muestran cómo cada método estructura relaciones de estilo en subconjuntos controlados.

La Tabla 2 y la Figura 1 revelan tendencias consistentes. *CSD* alcanza la mejor mAP@5 media (**0,59**), seguido de *CLIP-I* (0,55). En E_{avg}, *ALADIN* y *GOYA* logran la mejor media (0,21), con *CSD* cerca (0,22), mostrando que recuperación top-k y geometría global capturan aspectos complementarios. Aun así, ningún método actúa todavía como aproximación universalmente robusta

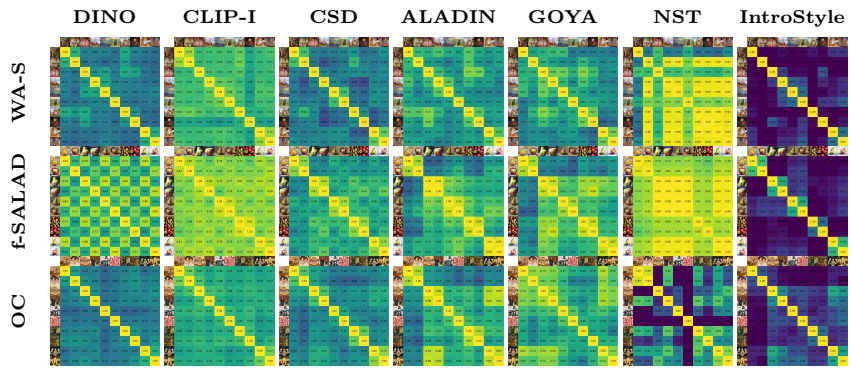


Fig. 1: Matrices de similitud de estilo en tres conjuntos de datos. Bloques diagonales claros indican coherencia intra-estilo; regiones oscuras fuera de la diagonal indican separación inter-estilo.

del juicio humano: el rendimiento varía notablemente entre conjuntos de datos y regímenes de etiquetas, especialmente en *SemArt*, donde todos obtienen baja mAP@5.

Los métodos de referencia genéricos *CLIP-I* y *DINO* suelen quedar por detrás en separación estilística, al estar optimizados para representación semántica o visual general. *CLIP-I* puede ser competitivo en varios conjuntos curados, sugiriendo que la recuperación se beneficia de correlaciones entre estilo y contenido recurrente. *DINO* rinde bien en algunos conjuntos de datos (mini-MET, Art500K), pero se degrada en otros. En cambio, métodos específicos como *ALADIN*, *CSD* y *GOYA* tienden a producir menor error y organización intra-estilo más coherente. *NST Gram* obtiene la menor mAP@5 media y el mayor E_{avg}, indicando que textura y paleta no bastan para discriminación robusta en dominios heterogéneos.

Los conjuntos de datos controlados, especialmente *fruit-SALAD*, son clave para interpretar el comportamiento: al desacoplar estilo y concepto, reducen atajos semánticos. Aquí, *ALADIN* y *CSD* destacan; *IntroStyle* logra alta recuperación, pero con mayor E_{avg}, sugiriendo una geometría global menos alineada. El rendimiento más débil de *DINO*, *CLIP-I* y *NST Gram* indica sensibilidad parcial a objeto, composición u otras pistas no estilísticas.

En conjunto, las métricas actuales aún no alcanzan una similitud de estilo de grano fino alineada con el juicio humano. Persisten dos limitaciones: separación incompleta estilo–contenido y robustez limitada entre dominios. Las afirmaciones sobre "comprensión del estilo" deben contextualizarse por el régimen de evaluación: buen rendimiento en conjuntos de datos entrelazados no implica necesariamente similitud fiable de estilo de grano fino.



Fig. 2: Ejemplos visuales de cuatro conjuntos de evaluación, mostrando dos estilos por conjunto y dos imágenes por estilo. Las parejas de imágenes en una misma fila permiten valorar la consistencia intra-estilo, mientras que la comparación entre filas dentro de cada columna ilustra la distinción inter-estilo.

6 Limitaciones de los conjuntos de evaluación actuales

Los resultados muestran que el rendimiento no depende solo del método, sino también de cómo cada conjunto de evaluación define "estilo". Algunos conjuntos de datos resultan más fáciles o difíciles para la mayoría de métodos, sugiriendo etiquetas heterogéneas: transformaciones visuales controladas, movimientos artísticos, identidad de artista, dominios amplios o metadatos museísticos. Así, una misma puntuación puede reflejar sensibilidad estilística, pero también temas recurrentes, tipo de objeto, material, autoría o duplicados.

Una segunda limitación es la correlación entre estilo y contenido. Muchos conjuntos curados no fueron diseñados para separar explícitamente ambos factores, por lo que las etiquetas de estilo pueden estar asociadas a motivos, géneros, composiciones u objetos concretos. Esto puede favorecer a métodos que capturan contenido semántico más que estilo visual. En este sentido, *fruit-SALAD* es especialmente informativo porque recombina explícitamente estilo y contenido, proporcionando una evaluación más directa de la invariancia al contenido.

Una tercera limitación es el soporte desigual por etiqueta. Muchos conjuntos reales presentan distribuciones desbalanceadas, con algunas clases muy representadas y muchas otras con pocas imágenes. Cuando una etiqueta de estilo cuenta con pocos ejemplos, resulta difícil estimar la consistencia intra-estilo y la tarea de recuperación puede depender de similitud a nivel de instancia más que de una noción estable de estilo.

Finalmente, los conjuntos también difieren en consistencia intra-estilo y distinción inter-estilo. Algunas etiquetas agrupan imágenes visualmente coherentes y bien separadas de otros estilos, mientras que otras cubren regímenes visuales amplios o heterogéneos. Los conjuntos controlados como *fruit-SALAD* y *OmniConsistency* suelen ofrecer etiquetas más definidas, mientras que las colecciones reales pueden contener categorías amplias o correlaciones sujeto–artista. La Figura 2 ilustra este efecto.

En general, estos conjuntos de evaluación no deberían interpretarse como intercambiables. Los controlados ofrecen evidencia más fuerte sobre consistencia de estilo y desentrelazamiento estilo–contenido; las colecciones grandes curadas o débilmente supervisadas son útiles para evaluar robustez bajo ruido de etiquetas y variabilidad real. El diseño de futuros conjuntos de evaluación debería priorizar control de contenido, soporte suficiente por estilo, manejo de duplicados y definiciones estilísticas más claras. Asimismo, futuras evaluaciones deberían incorporar alineamiento con juicios humanos y análisis de significancia estadística, especialmente ante sesgos y ruido de anotación en colecciones reales.

7 Conclusión

Este artículo revisó la similitud de estilo visual de grano fino, examinando conjuntos de datos, regímenes de evaluación y métodos, desde aproximaciones perceptuales genéricas hasta *embeddings* específicos de estilo. Además, realizamos una comparación unificada sobre conjuntos de evaluación diversos, usando recuperación y error de estructura de similitud para diagnosticar fuga de contenido y robustez entre dominios. Los resultados muestran que métodos dedicados como *CSD* superan consistentemente a métodos de referencia genéricos, aunque el rendimiento aún está lejos de una noción robusta y alineada con el juicio humano, especialmente en colecciones reales exigentes. Por tanto, las métricas fiables de estilo de grano fino siguen siendo un problema abierto. El trabajo futuro debería priorizar conjuntos de evaluación controlados, protocolos claros de desentrelazamiento y objetivos que impongan invariancia al contenido manteniendo sensibilidad a pistas estilísticas sutiles. También sería valioso estudiar consensos entre métricas, por ejemplo mediante agregación de similitudes o combinaciones de modelos, y utilizar desacuerdos entre métodos para detectar anotaciones ruidosas o potencialmente incorrectas.

Acknowledgments. Este trabajo está financiado por una ayuda predoctoral del Servicio de Investigación de la Universidad Pública de Navarra (contrato predoctoral de David Erroz) y por una ayuda FPU concedida a Rubén Pascual (ref. FPU22/02961), otorgada por el Ministerio de Ciencia, Innovación y Universidades de España (MICIU). También está financiado por el MICIU bajo los proyectos PID2022-136627NB-I00 (MCIN/AEI/10.13039/501100011033/FEDER, UE) y PID2025-174319OB-I00 (MICIU/AEI/10.13039/501100011033) y por el Gobierno de Navarra bajo el proyecto PC24-GENESIS-041-002.

References

1. Agarwal, D.: Artistic styles dataset. *Kaggle* dataset. `https://www.kaggle.com/datasets/devanshagarwal9/artistic-styles` (2023), accedido el 12 de julio de 2026
2. Art Dataset: Wiki Art Dataset. *Roboflow Universe* dataset. `https://universe.roboflow.com/art-dataset/wiki-art` (Mar 2022), accedido el 12 de julio de 2026

3. Bianco, S., Mazzini, D., Napoletano, P., Schettini, R.: Multitask painting categorization by deep multibranch neural network. Journal of Electronic Imaging **28**(1), 013,016 (2019). https://doi.org/10.1117/1.JEI.28.1.013016
4. Del Chiaro, R., Bagdanov, A.D., Del Bimbo, A.: Noisyart: A dataset for webly-supervised artwork recognition. In: Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019) – Volume 4: VISAPP. pp. 467–475. INSTICC, SciTePress (2019). https://doi.org/10.5220/0007392704670475
5. delayedkarma: Impressionist classifier data. *Kaggle* dataset. https://www.kaggle.com/datasets/delayedkarma/impressionist-classifier-data (2018), accedido el 12 de julio de 2026
6. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(5), 2567–2581 (2022). https://doi.org/10.1109/TPAMI.2020.3045810
7. Garcia, N., Vogiatzis, G.: How to read paintings: Semantic art understanding with multi-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (VISART) (2018)
8. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
9. Hessel, J., Holtzman, A., Forbes, M., Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2021)
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) (2017)
11. ikarus777: Best artworks of all time. *Kaggle* dataset. https://www.kaggle.com/datasets/ikarus777/best-artworks-of-all-time (2019), accedido el 12 de julio de 2026
12. Jin, X., Qiao, Q., Lu, Y., Wang, H., Huang, H., Gao, S., Liu, J., Li, R.: Apddv2: Aesthetics of paintings and drawings dataset with artist labeled scores and comments. In: Advances in Neural Information Processing Systems 37 (NeurIPS 2024). pp. 103064–103075 (2024). https://doi.org/10.52202/079017-3274
13. Kaggle: Painter by numbers. *Kaggle* competition. https://www.kaggle.com/c/painter-by-numbers (2016), accedido el 12 de julio de 2026
14. Khan, F.S., Beigpour, S., van de Weijer, J., Felsberg, M.: Painting-91: A large scale database for computational painting categorization. Machine Vision and Applications **25**(6), 1385–1397 (2014). https://doi.org/10.1007/s00138-014-0621-6
15. Khan, S., van Noord, N.: Stylistic multi-task analysis of ukiyo-e woodblock prints. In: Proceedings of the British Machine Vision Conference (BMVC). pp. 410.1–410.13 (2021). https://doi.org/10.5244/C.35.410
16. Koniusz, P., Schulz, R., Simmons, M., Stretcu, O., O'Brien, L.: Openmic: A large-scale open museum identification challenge dataset. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2018)
17. Kumar, A., Mu, J., Vasconcelos, N.: Introstyle: Training-free introspective style attribution using diffusion features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
18. Liao, P., Li, X., Liu, X., Keutzer, K.: The artbench dataset: Benchmarking generative models with artworks. arXiv preprint arXiv:2206.11404 (2022)

19. Ma, Z., Zhang, Y., You, Z., Tian, L., Liu, X.: Wikistyle+: A multimodal approach to content–style representation disentanglement for artistic image stylization. arXiv preprint arXiv:2412.14496 (2024)
20. Maerten, A.S., Chen, L.W., De Winter, S., Bossens, C., Wagemans, J.: Lapis: A novel dataset for personalized image aesthetic assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 6356–6365 (Jun 2025)
21. Mao, H., Cheung, M., She, J.: DeepArt: Learning joint representations of visual arts. In: Proceedings of the ACM International Conference on Multimedia (2017). https://doi.org/10.1145/3123266.3123405
22. Mensink, T., van Gemert, J.: The rijksmuseum challenge: Museum-centered visual recognition. In: Proceedings of International Conference on Multimedia Retrieval. p. 451–454. ICMR '14, Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2578726.2578791
23. Ohm, T., Karjus, A., Tamm, M.V., Schich, M.: fruit-salad: A style aligned artwork dataset to reveal similarity perception in image embeddings. Scientific Data **12**(1), 254 (2025). https://doi.org/10.1038/s41597-025-04529-4
24. Peng, X., Bai, X., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
25. programmer3: Artistic painting dataset. *Kaggle* dataset. https://www.kaggle.com/datasets/programmer3/artistic-painting-dataset, accedido el 12 de julio de 2026
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763 (18–24 Jul 2021)
27. Razadi, S.: WikiArt art movements/styles dataset. *Kaggle* dataset. https://www.kaggle.com/datasets/sivarazadi/wikiart-art-movementsstyles/data (2022), accedido el 12 de julio de 2026
28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10674–10685 (2022). https://doi.org/10.1109/CVPR52688.2022.01042
29. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 22500–22510 (2023). https://doi.org/10.1109/CVPR52729.2023.02155
30. Ruta, D., Gilbert, A., Aggarwal, P., Marri, N., Kale, A., Briggs, J., Speed, C., Jin, H., Faieta, B., Filipkowski, A., Lin, Z., Collomosse, J.: Stylebabel: Artistic style tagging and captioning. In: Computer Vision – ECCV 2022 Workshops, pp. 213–231. Springer (2022). https://doi.org/10.1007/978-3-031-20074-8_13
31. Ruta, D., Motiian, S., Faieta, B., Lin, Z., Jin, H., Filipkowski, A., Gilbert, A., Collomosse, J.: ALADIN: All layer adaptive instance normalization for fine-grained style similarity. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
32. Saleh, B., Elgammal, A.: Large-scale classification of fine-art paintings: Learning the right metric on the right feature. International Journal for Digital Art History **No 2** (2016). https://doi.org/10.11588/dah.2016.2.23376

33. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 25278–25294 (2022)
34. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3 (2025)
35. Somepalli, G., Gupta, A., Gupta, K., Palta, S., Goldblum, M., Geiping, J., Shrivastava, A., Goldstein, T.: Investigating style similarity in diffusion models. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024, vol. 15124, pp. 143–160. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-72848-8_9, series Title: Lecture Notes in Computer Science
36. Song, Y., Liu, C., Shou, M.Z.: Omniconsistency: Learning style-agnostic consistency from paired stylization data. In: NeurIPS 2025 (2025)
37. Strezoski, G., Worring, M.: Omniart: A large-scale artistic benchmark. ACM Transactions on Multimedia Computing, Communications and Applications **14**(4), 88:1–88:21 (2018). https://doi.org/10.1145/3273022
38. Tukur, M., Jashari, S., Hassanain, D.A., Bettio, F., Schneider, J., Pintore, G., Gobbetti, E., Agus, M.: PanoStyleVR: Style-based similarity metrics for web-based immersive panoramic style transfer. In: Proceedings of the 30th International Conference on 3D Web Technology (Web3D). pp. 1–10 (2025). https://doi.org/10.1145/3746237.3746287
39. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
40. Wilber, M.J., Fang, C., Jin, H., Hertzmann, A., Collomosse, J., Belongie, S.: Bam! the behance artistic media dataset for recognition beyond photography. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1211–1220 (2017). https://doi.org/10.1109/ICCV.2017.136
41. Wu, Y., Nakashima, Y., Garcia, N.: Goya: Leveraging generative art for content–style disentanglement. Journal of Imaging **10**(7), 156 (2024). https://doi.org/10.3390/jimaging10070156
42. Yaniv, J., Newman, Y., Shamir, A.: The face of art: Landmark detection and geometric style in portraits. ACM Transactions on Graphics **38**(4), 60:1–60:15 (2019). https://doi.org/10.1145/3306346.3322984
43. Ypsilantis, N.A., Garcia, N., Han, G., Ibrahimi, S., van Noord, N., Tolias, G.: The met dataset: Instance-level recognition for artworks. In: Vanschoren, J., Yeung, S. (eds.) Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. vol. 1. Curran (2021)
44. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
45. Zheng, M., Simsar, E., Yesiltepe, H., Tombari, F., Simon, J., Yanardag, P.: Stylebreeder: Exploring and democratizing artistic styles through text-to-image models. In: The Thirty-eight Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track (2024)