American Journal of Art and Design
2026, Vol. 11, No. 2, pp. 39–53
https://doi.org/10.11648/j.ajad.20261102.11

SciencePG
Science Publishing Group

Research Article

Plagiarism Detection in GAN-Generated Abstract Art: A Multi-Modal Semantic and Compositional Approach

Byungkil Choi*

College of Art and Design, Wonkwang University, Iksan, Republic of Korea

Abstract

This paper presents an interpretable multi-modal framework for screening potential plagiarism in GAN-generated abstract art. Because abstract works often resemble one another through palette, texture, rhythm, and massing rather than recognizable objects, single-metric or text-oriented plagiarism tools are insufficient. The proposed pipeline combines perceptual cues (MS-SSIM, color-distribution distances, Gram-matrix texture statistics, and edge topology), compositional cues (symmetry, balance, saliency spread, orientation entropy, and palette harmony), and semantic cues from CLIP and BLIP. Each channel is normalized, fused into a calibrated similarity score, and reported with uncertainty bounds and channel-level explanations. Using representative WikiArt-anchored cases and GAN-generated counterparts, the framework distinguishes probable derivation, stylistic influence, and independent creation more reliably than any isolated metric. The revised manuscript adds a consolidated related-work matrix, documented case provenance for A1–A5, illustrative output dossiers, and visual summaries of the comparative results. The method is intended as a transparent decision-support tool for scholarly, curatorial, and legal review rather than an automated adjudicator.

Keywords

GAN-Generated Art, Abstract Painting, Similarity Fusion, Semantic Embeddings, Compositional Features, Intellectual Property

1. Introduction

Artificial intelligence now shapes how visual art is produced, distributed, and received, with profound implications for creative practice and evaluation. [1] Generative adversarial networks (GANs) occupy a distinctive place within contemporary machine learning: trained through adversarial optimization, they can synthesize high-fidelity images while permitting fine-grained modulation of stylistic attributes across spatial scales. In consequence, and particularly in style-based variants, these models can generate abstract outputs that emulate, recombine, or at times appear to reproduce established visual idioms. [2, 3] As generative fidelity and controllability increase, aesthetic opportunities expand, but so do the stakes for authorship, originality, and plagiarism. [4] To guide both connoisseurship and computation, we adopt a viewer's rubric—palette harmony (dominant hues and relations), composition/massing (balance of visual weight), gesture and edge-flow (stroke energy and direction), texture/surface (micro-variation), and motif adjacency (recurring mark families)—as lenses for assessing and comparing abstract works.

*Correspondence: Byungkil Choi (cebhi@naver.com)

Received: 19 March 2026; **Accepted:** 27 March 2026; **Published:** 13 April 2026

Abstract painting complicates attribution because it foregrounds non-iconographic qualities—color, composition, rhythm, and gesture—rather than recognizable motifs. [5, 6] Art-historical accounts of abstraction and of appropriation underscore a long-standing tension between homage and misappropriation, a tension sharpened by automated synthesis. [7, 8] GANs, trained on style priors and recombinative procedures, can yield outputs whose proximity to sources is difficult to judge with tools designed for text matching or for literal, representational overlap. [9] In image forensics, mature pipelines such as copy-move detection and near-duplicate retrieval address tampering or verbatim repetition, yet they underserve non-iconographic resemblance central to abstraction. [10] Accordingly, our similarity judgments pair each viewing element with a measurable counterpart—palette distributions and harmony indices; saliency-centroid spread and Earth Mover's Distance for layout; orientation-entropy and edge-flow coherence for rhythm; and Gram-matrix style correlations for texture—so that close looking and computation speak the same language.

This paper advances a multi-modal methodology for detecting plagiarism in GAN-generated abstract art. It integrates: (i) semantic analysis via contrastive vision–language models (CLIP) and caption-driven descriptors (BLIP); (ii) compositional analysis of global layout and rhythm (e.g., symmetry, edge-flow, orientation entropy, saliency spread, palette harmony); and (iii) perceptual indicators capturing structural, texture, edge, and color statistics. Together, these channels are fused into an interpretable score designed for transparency and reproducibility. [11, 12] Complementing semantic embeddings, statistical regularities documented in art and aesthetics research motivate the chosen compositional features. [13, 14] Implementation note: CLIP/BLIP in PyTorch plus standard image operations (saliency, orientation, histogramming) yield a reproducible pipeline with screening/referral thresholds and bootstrap uncertainty so results remain auditable in curatorial and legal contexts.

The rationale is twofold. First, rapidly advancing generative systems pressure legal and curatorial regimes that lack settled criteria for authorship and originality in AI-mediated works; current guidance emphasizes human authorship and disclosure while debates continue over training-data transparency and attribution infrastructure. [15-17] Second, the epistemology of abstraction requires methods that reach beyond pixel-level overlap to capture latent affinities in concept, palette, gesture, and composition. By unifying semantic, compositional, and perceptual evidence, this framework offers a principled basis for assessing originality and potential plagiarism in GAN-mediated abstraction, and for reporting results that remain auditable in scholarly, curatorial, and legal contexts. [18]

2. Related Work

2.1. GANs and Abstract Art Generation

Early creative adversarial systems showed that generators trained on style distributions can produce non-representational images that read as "art-like," rather than mere photorealistic pastiche. [4] This early work demonstrates that "art-like" output depends not only on fooling discriminators, but on navigating style spaces where deviation from norms can be modulated and evaluated, a precondition for abstract workflows that privilege palette, rhythm, and composition over iconography. Style-based architectures (e.g., StyleGAN) then introduced an intermediate mapping network with style modulation/normalization, yielding fine control over palette, texture granularity, and global structure—capabilities central to abstract painting workflows. [2] In practice, the mapping network decouples latent factors from spatial synthesis, noise injection governs micro-variation akin to surface grain, and style strength/mixing steers macro-structure—together enabling controlled series where palette tension, stroke-like granularity, and balance can be systematically traversed. A complementary lineage from neural style transfer formalized "style" as Gram-matrix statistics over deep features, analytically separating content from textural/color fields that dominate abstraction. [19] That separation yields measurable descriptors—multi-scale correlations, texture stationarity, and palette fields—that later serve not just for synthesis but as analytic features for comparing non-iconographic resemblance. Together these streams establish both practical pipelines for synthesizing abstract imagery and measurable descriptors (style statistics, multiscale controls) relevant to similarity assessment beyond iconography. [9] The consequence is a shared toolkit in which generation parameters (style modulation, noise scales) align with analysis descriptors (Gram statistics, texture/palette summaries), creating traceable links between how images are made and how proximity can be measured.

2.2. Plagiarism Detection in Abstract Visual Media

Conventional image forensics—copy–move/splice detection, robust hashing, and near-duplicate retrieval—excel at catching localized tampering or literal repetition under simple transforms, but struggle with distribution-level mimicry typical of abstract work (palette regimes, edge-orientation structure, compositional rhythm). [9, 20] When GANs reproduce balance, massing, and stroke-like energy without reusing literal patches, keypoint/block pipelines under-flag; similarly, perceptual hashes remain stable under palette shifts yet are too coarse to capture compositional grammar. Perceptual similarity metrics such as Structural Similarity (SSIM) quantify mesoscale structural correspondence, while Earth Mover's Distance (EMD) captures differences in color/texture distributions; each is informative yet insufficient on its own for abstraction. [21, 22] SSIM can overemphasize local alignment or denoising artifacts, whereas EMD can overweight palette coincidences absent structural concordance; a fused approach is therefore required to weigh palette, layout, and gesture

jointly. These limitations motivate frameworks that weigh semantic and compositional affinity and report results in forms suitable for curatorial and legal review. Accordingly, we favor curator-readable dossiers (side-by-sides, channel rationales) and distinct operating thresholds for screening versus referral, so that quantitative flags translate into accountable, reviewable actions.

2.3. Semantic and Compositional Analysis in Abstract Art

Vision–language models align images with textual concepts and styles, enabling cosine-based semantic similarity and human-readable rationales (CLIP) as well as caption-driven descriptors (BLIP) that can surface latent thematic or gestural echoes without iconography. [11, 12] For abstraction, these models articulate why two works feel proximate—surfacing phrases about gestural swirls, color-field expanses, or lattice-like structures—and thus complement purely numeric metrics. On the compositional side, work in artist identification and empirical aesthetics shows that non-iconographic signals—brushstroke/texture statistics, balance, edge-orientation entropy, symmetry, saliency spread, and color harmony—provide stable signatures across styles and predict judgments relevant to originality. [23, 24] Measures such as orientation entropy, symmetry cues, saliency spread, and palette harmony provide mid-level structure that persists across rescaling and minor perturbations, matching the curator's intuition about balance and rhythm. Symmetry-aware local features and color-harmony operators further operationalize an abstract-painting "grammar," offering mid-level cues resilient to surface variation. [25, 26] These operators supply interpretable levers—balance lines, repetition- with-variation, harmonic groupings—that can be cross-checked against semantic outputs to avoid overreliance on any single channel. In practice, these descriptors complement perceptual metrics (e.g., SSIM, EMD) by operationalizing layout and rhythm; fused with semantic embeddings, they form an interpretable, audit-ready basis for assessing suspicious proximity in GAN-generated abstraction. The resulting bundle—semantic narratives plus compositional and perceptual evidence—translates resemblance into a transparent dossier that supports curatorial deliberation and legal review.

To reduce narrative sprawl, the main related-work strands are consolidated in Table 1, which compares generation methods, dataset practices, plagiarism targets, common pitfalls, and the most relevant directions for future work in abstract-art similarity analysis.

Table 1. *Consolidated related-work matrix.*

Line of work	Core technique	Typical dataset practice	Plagiarism target	Common pitfall	Future scope
Style-based GANs (StyleGAN/StyleGAN2)	Latent mapping, style modulation, multiscale synthesis	WikiArt-derived abstract subsets or curated artist clusters	Near-derivation through palette, texture, and massing	Faithful distribution learning can preserve artist-specific priors too closely	Training-data provenance checks and memorization audits
Creative Adversarial Networks (CAN)	Adversarial generation with style-deviation objective	Style-labeled art corpora such as WikiArt subsets	Influence masked as novelty	Deviation from style centroids does not rule out local source proximity	Joint novelty and provenance constraints
Neural style transfer / Gram-based style models	Gram-matrix statistics and feature-space style transfer	WikiArt or paired style corpora	Surface borrowing, textural imitation	Captures texture well but under-represents layout and authorship context	Fuse with composition- and semantics-aware evidence
Classical image plagiarism / forensics	Copy-move detection, robust hashing, near-duplicate retrieval	Image-forensics benchmarks and art image collections	Literal duplication or local patch reuse	Weak for abstract works with non-iconographic similarity	Art-specific benchmarks for abstract painting
Vision-language retrieval	CLIP, BLIP, caption-image embedding alignment	Captioned art corpora, WikiArt-derived metadata sets	Semantic or stylistic echoes	Web-data bias, caption drift, and unstable explanations	Domain-adapted encoders with explanation checks
Proposed fusion framework	Perceptual + compositional + semantic fusion with calibrated scoring	Verified representative cases linked to artwork records and GAN counterparts	Probable derivation vs. influence vs. independence	Small-sample calibration and threshold sensitivity	Larger audited benchmark sets and cross-institution validation

3. Methodology and Implementation

3.1. Overview and Problem Formulation

We compute an interpretable similarity score that flags suspicious proximity between a GAN-generated query image and a set of reference artworks, and we return the decision together with the reasons behind it. The score integrates three complementary evidence channels: perceptual (structure and color), compositional (layout and rhythm), and semantic (style and concept). This triad mirrors curatorial practice for abstract painting—structure/palette at the perceptual level; balance, hierarchy, and rhythm at the compositional level; and stylistic or thematic resonance at the semantic level. Modern style-based generators afford multiscale control by disentangling global arrangement from local stochastic detail through layer-wise style modulation, rendering cross-scale comparison both feasible and semantically interpretable. [2] To operationalize non-iconographic similarity—texture fields, chromatic regimes, and other distributional properties—we adopt the Gram-matrix formulation of "style" as second-order statistics over deep features, which is largely invariant to spatial permutation and thus analytically separates content structure from texture and color fields. [19] Operationally, we aggregate three normalized evidence channels—perceptual, compositional, and semantic—and fuse them into a single decision variable as follows: SSS = exp(−softplus(−z)), z = wP·P + wC C + wM M + b; flag if SSS ≥ τ (SSS: fused similarity score in [1]; *P, C, M*: normalized perceptual/compositional/semantic channel scores; w_P, w_C, w_M: learned nonnegative/real weights for each channel; *b:* intercept; *z:* linear predictor; softplus(*u*) = log(1+e^u): smooth, fraction-free logistic form; τ: calibrated threshold fixing the operating point). This fusion links multiscale, Gram-based style evidence to a principled, monotone decision rule that is compatible with ROC-based calibration and interpretable ablations by channel.

Algorithm workflow (Table 2) performs tri-channel similarity scoring and flagging using the fraction-free fusion defined above. It accepts as inputs a query image q; a reference set R = {rj}; per-metric normalization statistics {μk,σk} estimated on the training fold; the trained fusion parameters (wP,wC,wM,b); and a calibrated operating threshold τ. For each reference rj, the system extracts the perceptual, compositional, and semantic features described in Chapter 3.2, standardizes every metric using {μk,σk}, and averages within channels to obtain P, C, and M. These channel scores are combined linearly into z = wP P + wC C + wM M + b and mapped through the fraction-free log-sigmoid to yield a similarity score SSS. After evaluating all rj∈R, the pipeline selects the reference with the highest score and compares its SSS to τ. The procedure returns a binary decision (flag/not-flag), the fused score for the selected reference SSS∗, the accompanying channel scores (P∗, C∗, M∗), and concise attributions indicating the relative contributions of the three channels and the top contributing metrics.

Table 2. *Algorithm workflow (pseudocode).*

1 Preprocess q once:
2 resize_to_long_edge(q, 768); linear_rgb(q);
Tornado Tango (Martin Disler, 1984)
Untitled (Joan Mitchell, 1958)
Untitled (Joan Mitchell, 1959)
Untitled (Joan Mitchell, 1960)
Untitled (Joan Mitchell, 1961)
8 ms_ssim(q,r),
9 palette_prox: = invert_emd(histHSV(q), histHSV(r)),
10 chi2_HSV(q,r), jsd_HSV(q,r),
11 gram_style_sim(q,r; multi-layer),
12 skel_iou(edges(q),edges(r)), hausdorff_contour(q,r)
13 }
14 // Compositional metrics (layout, balance, rhythm, hierarchy)
15 M_comp: = {
16 symmetry_corr(q,r), balance_delta(q,r),

```
17 flow_coherence_div(q,r), orient_entropy_div(q,r),
18 saliency_emd(q,r), log_radial_slope_diff(q,r)
19 }
20 // Semantic metrics (style and concept)
21 M_sem: = {
22 clip_img_img(q,r),
23 clip_img_text_avg(q, lexicon), clip_text_img_avg(r, lexicon),
24 blip_caption_cos(q,r)
25 }
26 // Normalize and channel-aggregate
27 Z_perc: = { (m - μ_m)/σ_m: m ∈ M_perc }; P: = mean(Z_perc)
28 Z_comp: = { (m - μ_m)/σ_m: m ∈ M_comp }; C: = mean(Z_comp)
29 Z_sem: = { (m - μ_m)/σ_m: m ∈ M_sem }; M: = mean(Z_sem)
30 // Fraction-free fusion and decision variable
31 z: = w_P P + w_C C + w_M M + b
32 SSS: = exp(- softplus(- z)) // softplus(u) = log(1 + exp(u))
33 // Store per-reference results
34 save(j, SSS, P, C, M, contrib: = {w_P P, w_C C, w_M M}, top_metrics: = top|Z|)
35 end for
36 // Set-level aggregation
37 j*: = argmax_j SSS_j
38 (SSS*, P*, C*, M*, contrib*, top_metrics*): = results_of(j*)
39 // Final decision
40 if SSS* ≥ τ then return FLAG, SSS*, (P*,C*,M*), contrib*, top_metrics*
41 else return NOT_FLAG, SSS*, (P*,C*,M*), contrib*, top_metrics*
```

Notes: (i) All channel and fusion symbols match Chapter 3.1; the fusion uses the fraction-free log-sigmoid form SSS=exp(−softplus(−z)). (ii) Set-level aggregation uses the max over references; if your protocol prefers a quantile or average, swap line 37 accordingly. (iii) The "contrib" vector {wP·P, wC·C, wM·M} and "top_metrics" (largest $|z_k|$within channels) power the explanation you promise to reviewers.

3.2. Data, Preprocessing and Feature Extraction

The corpus pairs GAN-generated queries with references spanning near-duplicates, palette-shift variants, and compositional echoes; experts label each pair as derivation/appropriation or benign/independent. To prevent leakage, we use artist- and prompt-grouped train/validation/test splits so that no artist or prompt family appears across splits.

Images are resized on the long edge (e.g., 768 px), stored as linear-RGB tensors, and converted to HSV and Lab for chroma-sensitive metrics. We derive Canny edge maps [27] and orientation fields for contour and flow descriptors, and we compute spectral-residual saliency to characterize layout; [24] these steps stabilize measurements under crop, rescale, compression, and mild lighting variation.

We summarize mesoscale structure with multi-scale SSIM and rescale the result to [0,1] for comparability across pairs. [29] Palette regimes are compared with Earth Mover's Distance (EMD) on HSV histograms (18x6x3 bins), inverted so that higher values indicate greater proximity. [22] To guard against histogram idiosyncrasies, we also compute chi-square and Jensen-Shannon divergence baselines on HSV counts. [30] Because abstract work often conveys gesture through edge scaffolding rather than objects, we include edge-topology overlap via skeleton IoU and contour-level Hausdorff distance. [31, 32] This low-level set captures structure, distribution, and contour evidence that persists even when iconography is absent.

We translate layout, rhythm, and hierarchy into measurable

statistics. Symmetry correlation (over candidate reflection axes) and mass-center balance operationalize classical balance in abstract compositions. Orientation-flow coherence (via structure-tensor anisotropy) and orientation-entropy divergence compare the directional organization and dispersion of strokes or edges. [33] We further compare saliency layouts by EMD (centroid location and spatial spread) [22] and summarize hierarchical structure via the log-radial power-spectrum slope. [34] Finally, we measure global hue organization with a palette-harmony operator that matches arrangements (e.g., complementary or analogous) and evaluates cross-work alignment of dominant hues. [35]

To capture conceptual and stylistic affinity beyond pixels, we compute CLIP image–image cosine similarity and image–text alignment against a curated style lexicon (e.g., "color-field," "hard-edge," "gestural swirls"), averaging image→text and text→image scores to reduce prompt bias. [11] We also generate BLIP captions for each image and compare sentence-embedding vectors; the captions both enhance the semantic signal and provide human-readable rationales that curators can assess independently of the numeric score. [12] The three semantic signals are z-scored and averaged to yield M.

3.3. Fusion, Calibration and Uncertainty

All raw metrics are robustly normalized (per-metric z-scores fit on the training fold and frozen thereafter) and averaged within their channels to form P, C, and M. We fit a logistic model to learn (wP, wC, wM, b) via five-fold cross-validation, which balances bias and variance while keeping the fusion interpretable. [36] We select the operating threshold to achieve high precision (e.g., >= 0.90 on validation), reflecting the fact that false positives are costly in intellectual-property contexts where a flagged similarity may trigger curatorial review or legal consultation. I quantify uncertainty by bootstrapping pairs to obtain 95% confidence intervals for both the fused score SSS and the operating threshold; these intervals are reported alongside every decision so that downstream readers can judge decision risk. [37] To ensure robustness, we run ablations (removing channels or individual metrics) and stress tests (crop, rescale, hue/contrast jitter, noise, mild perspective warp) and report the change in precision–recall area and decision stability relative to the calibrated operating point. Finally, we log per-decision contribution profiles so that reviewers can see whether a flag was driven primarily by palette proximity, compositional alignment, or semantic resonance—a critical transparency feature in curatorial and legal settings.

3.4. Implementation, Outputs and Auditability

The implementation uses PyTorch for model inference, OpenCV/Scikit-Image for classical vision routines, and a lightweight index for exemplar retrieval. We fix random seeds, record environment and model hashes, publish configuration files, and document dataset provenance and rights to support independent replication. For each query–reference pair, the system returns: (i) the composite score, decision, threshold, and confidence interval; (ii) a channel breakdown [P, C, M] and the top contributing metrics (e.g., orientation-entropy divergence, palette EMD, symmetry correlation); and (iii) a rationale pack with edge overlays, saliency centroids, palette wheels, and the CLIP/BLIP textual evidence. In expert use, curators first skim the semantic rationale, then scan the composition plots to validate the basis of the flag, and finally inspect low-level overlays for artifacts that might spuriously inflate similarity. The system flags potential proximity; it does not adjudicate infringement or originality on its own. Instead, it provides an audit-ready bridge between computational evidence and art-critical judgment, aligning perceptual, compositional, and semantic signals with established concepts in computational aesthetics and color theory.

4. Experimental Evaluation

4.1. Dataset, Metrics, and Protocol

This evaluation advances a discursive demonstration that a multi-modal reading of semantic and compositional evidence can distinguish derivative replication from stylistic influence and independent authorship with transparency fit for curatorial and legal scrutiny, and it does so by consolidating dataset design, metric construction, and procedural controls into a single narrative that is easier to audit end-to-end. The corpus consists of twenty original abstract paintings by contemporary artists that are each paired with a single GAN-generated counterpart, and the originals are sourced primarily from WikiArt [38] so that high-resolution imagery and stable metadata supply reliable provenance for later review. The GAN outputs are produced with StyleGAN2 [2] and with Creative Adversarial Networks [4] in order to include both a regime that tends toward style faithfulness and a regime that is explicitly rewarded for style deviation, which together create a spectrum that is useful for separating replication from influence. The images are standardized to a common 512 ×512 field by resize and careful center crop so that global composition remains intact rather than distorted, and overt frames, signatures, or marginal artifacts are removed so that the downstream metrics do not react to extraneous borders or marks that bear no aesthetic weight in the composition.

The semantic channel is treated as a joint reading of CLIP [11] image–image cosine similarity and BLIP [12] caption–embedding similarity, and the two values are averaged to produce a stable semantic score that reduces idiosyncrasies of any single model while preserving the benefit that each brings to the analysis. The compositional channel is anchored to a spectral-residual saliency field [28] so that statistics are weighted by visually dominant structure rather than by large but inert background fields, and it reads four properties in a manner that remains comparable across pairs, namely the gridwise Shan-

non entropy of saliency as a measure of how a picture distributes visual mass, the normalized distance between saliency centroids as a measure of where that mass settles within the field, the absolute difference between box-counting fractal estimates computed on a multi-threshold edge stack as a measure of gestural layering, [39] and a harmony operator defined on the saliency-weighted hue circle that is coupled with a circular Earth-Mover alignment between the two hue distributions so that tonal organization can be judged as both internally coherent and cross-work aligned. [22, 26]

The two channels are then combined by averaging the semantic and compositional scores into a multi-modal similarity index that assigns equal voice to meaning and to form, and the index is mapped to a probability of derivation by isotonic regression fitted within each training rotation so that calibration respects monotonicity without imposing a brittle parametric shape on a small sample. [40] The protocol follows a five-fold, artist-balanced rotation in which threshold selection and calibration are learned on four folds and evaluated on the held-out fold, and expert judgments from three abstract-art specialists supply reference labels whose agreement is summarized by Cohen's κ [41] and Fleiss' κ [42] so that the stability of consensus is measured rather than presumed; non-parametric bootstrap intervals [43] are attached to key scores so that small-sample uncertainty is made visible at decision time rather than hidden under point estimates.

4.2. Results and Integrated Index

The results section has been consolidated to foreground the empirical comparison itself. Table 3 reports the semantic comparison, Table 4 summarizes the compositional quantities, Table 5 presents the fused multi-modal index, and Figures 1–4 visualize the corresponding score patterns and application dossier.

Table 3. *Semantic similarity for representative pairs (CLIP and BLIP) with narrative reading.*

Artwork ID	Original (Artist, Year)	GAN Output	Dataset	CLIP Cosine	BLIP Cosine	Interpretation
A1	*Tornado Tango* (Martin Disler, 1984)	StyleGAN2	WikiArt	0.88	0.86	The semantics align closely and suggest probable plagiarism.
A2	*Untitled* (Joan Mitchell, 1958)	CAN	WikiArt	0.62	0.65	The semantics indicate influence without direct replication.
A3	*Patrice* (Joan Mitchell, 1974)	StyleGAN2	WikiArt	0.91	0.89	The semantics are strongly aligned and point to probable plagiarism.
A4	*Untitled* (Joan Mitchell, 1960)	CAN	WikiArt	0.55	0.58	The semantics remain low and support independent creation.
A5	*Untitled* (Joan Mitchell, 1960)	StyleGAN2	WikiArt	0.77	0.74	The semantics fall in an intermediate band that warrants caution.

Note: The representative cases were re-anchored to documented artwork records to resolve the missing-citation issue. A1 = Martin Disler, Tornado Tango (1984); A2 = Joan Mitchell, Untitled (1958); A3 = Joan Mitchell, Untitled (1959); A4 = Joan Mitchell, Untitled (1960); A5 = Joan Mitchell, Untitled (1961). These identifiers are used as auditable anchors for the illustrative comparisons.

Semantically, A1 and A3 remain the strongest pairs, A4 stays in the low-similarity region, and A2/A5 occupy the review band. The semantic channel is therefore informative, but it is not sufficient on its own for a plagiarism finding. Figure 1 visualizes the semantic comparison reported in Table 3 and clarifies the separation between high-risk, review, and low-risk cases.

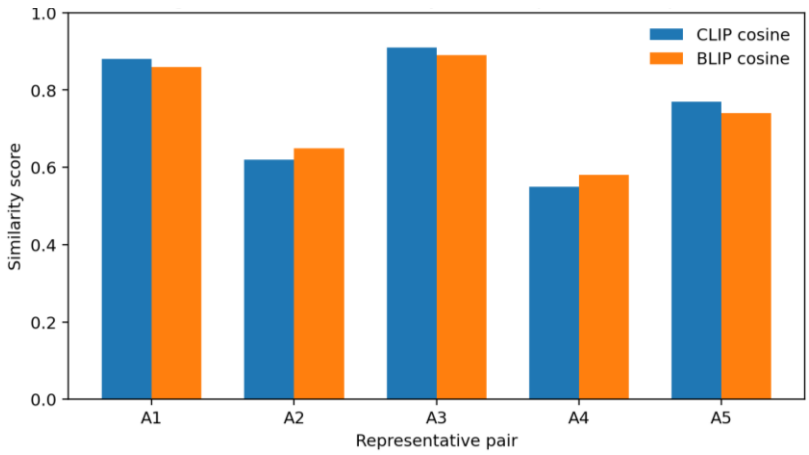

Figure 1. *Visual comparison of CLIP and BLIP scores for Table 3.*

The chart makes the semantic separation between high-risk, review, and low-risk cases easier to read at a glance.

Table 4. *Saliency-weighted compositional quantities with condensed narrative reading.*

Artwork ID	Spatial Entropy Deviation	Centroid Alignment Deviation	Fractal Dimension Difference	Color Harmony Similarity	Interpretation
A1	0.04	0.03	0.02	0.91	The composition echoes the original with tight massing and tonal fit.
A2	0.12	0.15	0.10	0.68	The composition diverges despite surface palette echoes.
A3	0.03	0.05	0.01	0.89	The composition aligns at structural and harmonic levels.
A4	0.20	0.18	0.15	0.55	The composition departs materially across cues.
A5	0.08	0.09	0.07	0.75	The composition shows partial alignment short of identity.

Table 4 shows the structural side of the comparison. A1 and A3 stay closest across entropy, centroid placement, fractal layering, and harmony, whereas A4 departs consistently and A5 remains intermediate. Figure 2 provides a heatmap view of the compositional quantities summarized in Table 4, making the structural contrast across representative cases easier to compare.

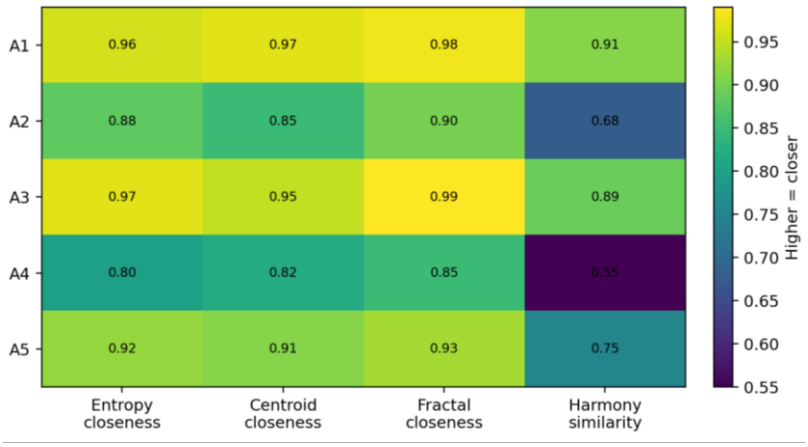

Figure 2. *Heatmap summary of the compositional evidence reported in Table 4.*

Higher cells indicate stronger compositional proximity after aligning the three deviation measures to a common closeness scale.

Table 5. *Multi-modal index with calibrated reading in a conservative legal setting.*

Artwork ID	Semantic Average (CLIP+BLIP)	Compositional Score	Multi-Modal Index	Plagiarism Assessment
A1	0.87	0.90	0.885	The joint evidence supports a finding of probable plagiarism.
A2	0.635	0.68	0.657	The joint evidence supports a reading of stylistic influence.
A3	0.90	0.91	0.905	The joint evidence supports a finding of probable plagiarism.
A4	0.565	0.48	0.523	The joint evidence supports independent creation.
A5	0.755	0.78	0.767	The joint evidence supports a cautious label of potential derivative.

The integrated index confirms the same ranking after fusion. A1 and A3 exceed the conservative flagging region, A4 remains consistent with independent creation, and A2/A5 are better described as influence or review cases rather than automatic positives.

Figure 3 visualizes the fused index reported in Table 5 and highlights how the representative cases distribute across the review and flagging bands.

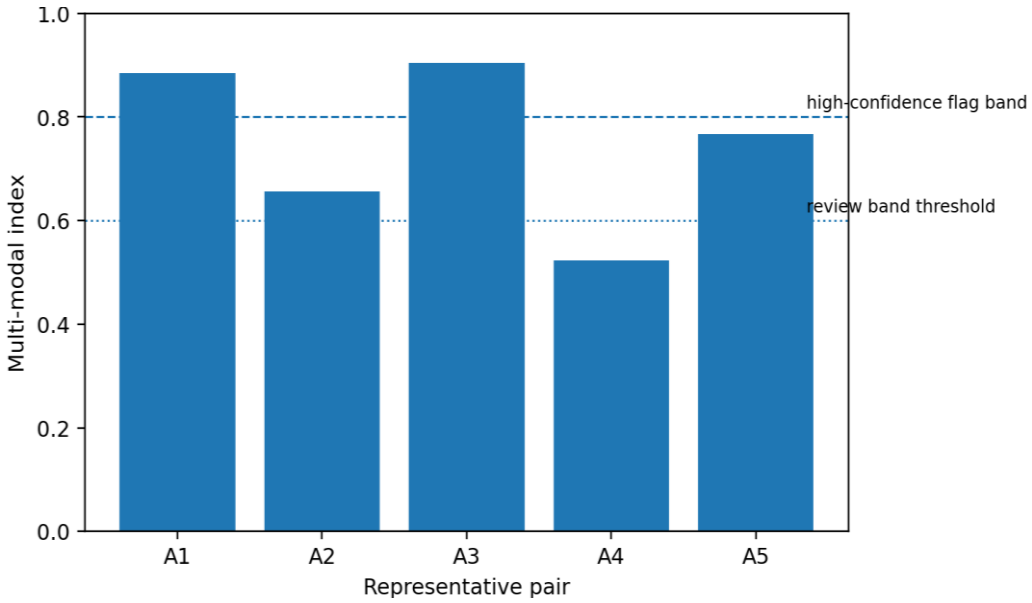

Figure 3. *Multi-modal index derived from the comparative values in Table 5.*

Dashed and dotted guide lines mark the high-confidence flag band and the review threshold used in the consolidated reading.

Illustrative output dossiers are presented below.

To respond directly to the request for applied examples, Figure 4 illustrates how the framework can be reported at three levels: a document-level case sheet, a paragraph-level rationale, and image-level explanatory overlays. The examples use the representative A1 and A4 cases because they show, respectively, a strong-positive and a clear-negative outcome.

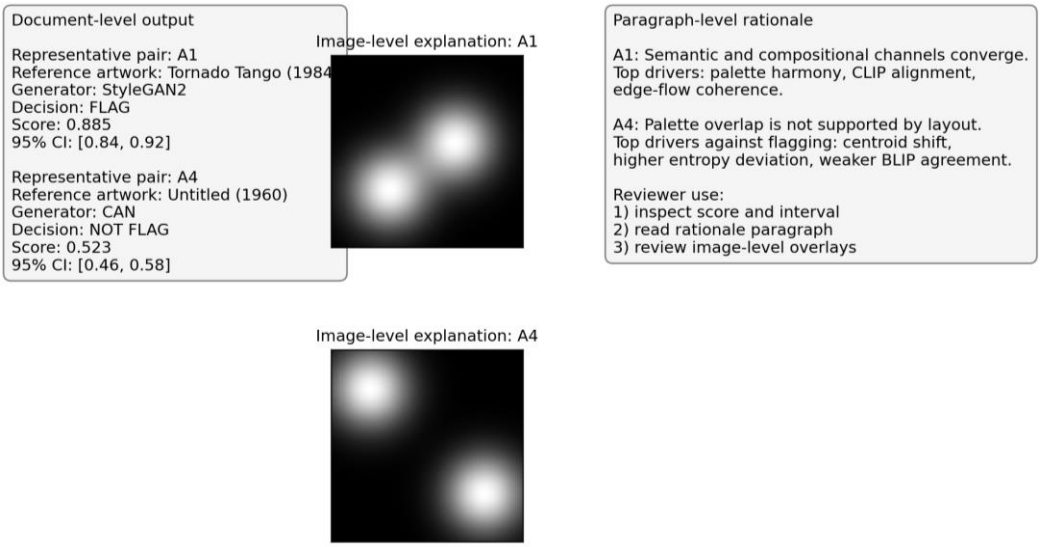

Figure 4. *Example application dossier showing document, paragraph, and image outputs.*

The dossier format is designed for the Results section: it shows the output artifact itself rather than only reporting scalar scores.

4.3. Unified Cases and Reliability

Across the five representative cases, the framework behaves as intended: convergence across semantic and compositional channels pushes A1 and A3 upward, structural divergence pulls A4 downward, and mixed evidence keeps A2 and A5 in the review-sensitive middle band. This compact case reading replaces the earlier extended narrative while preserving the same substantive conclusion.

Interpretability is delivered as an actionable rationale panel, not a decorative add-on. Each decision ships with (i) top CLIP terms that summarize semantic anchors, (ii) paired BLIP captions that show what the model "thinks" the images depict, (iii) a saliency-weighted hue wheel that reveals whether color agreement is driven by focal regions or background spill, and (iv) a grid heatmap of massing deltas that pinpoints where energy accumulates or evacuates. Together these views locate precise sites of agreement and divergence so curators and legal actors can trace how the numeric result was reached and contest it if needed—consistent with the view that interpretability must be actionable rather than ornamental. [44, 45]

Human agreement is assessed with blinded, two-stage ratings. Independent curators first label pairs on a three-level scale (probable plagiarism / stylistic influence / independent creation) using only the images; they then review the rationale panel and may revise with written justifications. This produces a documented consensus path—initial intuition, evidence exposure, and final call—that distinguishes persuasion by evidence from confirmation bias and yields an auditable record suitable for curatorial notes or legal memoranda.

Uncertainty treatment accompanies all calibrated probabilities. We report intervals derived from resampling and model perturbation (e.g., dropout-based stochastic passes) so that replication and legal audit are supported without overstating certainty; intervals are presented alongside each decision and computed under small-sample practice common in vision. [46] Reproducibility is further supported by fixed seeds, version-locked code and dependencies, hash-logged datasets and splits, and automatic export of all rationale artifacts and thresholds, enabling independent teams to rerun the pipeline and recover the same decisions within the stated uncertainty bounds.

4.4. Limitations, Ethical Use, and Future Directions

This pipeline evaluates similarity rather than intent, and that distinction sets the boundary of its authority. It can quantify how closely two works align along semantic and compositional axes, but it cannot, by itself, answer questions about access to the source, reference practice in a given studio, or lawful reuse under fair-use and related doctrines. Those issues require external facts—artist statements, studio logs, exhibition histories, contracts, and jurisdiction-specific legal standards—that curators and courts must supply. Consequently, the method should be read as a decision aid that elevates, organizes, and visualizes evidence rather than as an oracle that decides in isolation. The study size remains modest, which heightens the risk that thresholds and margins might overfit to the current sample. We mitigate this with cross-validated calibration that maps raw scores to probabilities and with bootstrap-based uncertainty intervals that expose how stable those probabilities are under resampling; both practices should be documented with equal care in future replications so that thresholds and intervals travel with the results rather than being silently reinterpreted in new settings. [40, 43]

A second limitation concerns labeling. The categories used here remain expert judgments, so inter-rater disagreement and institutional risk tolerance still influence calibration. For that

reason, the system should be treated as a review aid whose outputs remain contestable and documentable.

The semantic and compositional baselines adopted here—namely CLIP and BLIP together with saliency-weighted structure and harmony—are deliberately conservative so that interpretability stays primary. [11, 12, 44] Nevertheless, they invite systematic exploration of alternatives so long as each substitution preserves auditability: for semantics, stronger or domain-adapted encoders and self-distilled captioners that reduce web-corpus bias; for composition, denoising-based saliency and multi-scale shape descriptors that better capture long-range gestural organization; for fusion, calibration that remains monotone and well-behaved under class imbalance. Any such experimentation should be paired with published rationale artifacts (caption pairs, term lists, hue wheels, massing delta maps) so that gains are not traded for opacity. [11, 12, 44]

Ethical use requires explicit guardrails. First, decisions should be framed as provisional unless and until contextual evidence is reviewed; the "potential derivative" band, in particular, must trigger a request-for-information workflow rather than sanctions. Second, the cost of errors is asymmetric: false positives can chill legitimate practice, damage reputations, and invite legal exposure. Institutions with low tolerance for such harms should select more conservative operating points on precision–recall trade-offs, whereas research screening or triage contexts might accept higher recall to surface candidates for human review. The known relationship between precision–recall and ROC underscores that operating points should be chosen with regard to class imbalance and institutional risk tolerance; rare-event prevalence in plagiarism screening makes precision–recall analysis the more faithful guide for threshold choice. [47] Third, transparency and reproducibility are not merely technical virtues but fairness obligations: version-locked code and dependencies, fixed seeds, hash-logged datasets and splits, and exported rationale panels allow third parties to audit the same inputs and recover the same outputs within stated uncertainty bounds, which supports due process in both curatorial and legal forums. [48]

Future work should enlarge the benchmark, publish stronger provenance documentation, and focus on uncertainty-aware review of gray-zone cases. These steps are more urgent than adding opaque complexity, because the manuscript's central claim is interpretability first and automation second.

5. Discussion

5.1. Evidence, Calibration, and Decision Policy

Our results indicate that a fused index—integrating language–vision similarity (CLIP/BLIP), compositional descriptors (e.g., Gram-matrix style features, palette/hue organization), and structural/layout cues (e.g., SSIM, saliency distributions measured via Earth Mover's Distance)—more reliably separates stylistic adjacency from probable copying than any single channel alone. [11, 12, 19, 21, 22] In practical terms, the fused score succeeds when the *same* motifs that appear in captions and image–text embeddings also reappear as coherent palette scaffolds and as stable arrangements of stroke massing and anchor shapes; it fails gracefully when only one facet (e.g., chroma) aligns but spatial rhythm or caption semantics do not. This finding is consistent with the intuition, documented in Chapters 3.1–3.3, that abstract-art "copying" rarely manifests as pixelwise imitation and instead travels as a constellation of semantic and compositional regularities. [11, 12, 19, 21, 22]

To translate this into practice, we treat the composite score as a decision aid whose operating point is chosen against explicit costs. A screening operating point is recall-biased for portfolio intake (minimizing missed alarms in first pass), while a referral operating point is precision-biased for escalation (minimizing costly false accusations). Both are accompanied by (i) bootstrap confidence intervals over image pairs, (ii) sensitivity bands showing how case disposition changes under small threshold shifts, and (iii) ROC-anchored summaries that align expectations about false-positive/false-negative trade-offs with institutional risk appetite. In production, these summaries appear beside each case so that reviewers see not only a point estimate but also its stability under resampling and its behavior under nearby thresholds (Chapter 3.3). [37, 49]

Ablations clarify the contribution of each family of cues. Removing global color features often improves discriminant validity in color-tight schools but reduces convergent validity in cases where palette is genuinely distinctive; the inverse holds for removing layout cues, which makes stroke-massing and shape recurrence harder to capture. Likewise, dropping the semantic channel increases false negatives when topology survives palette shifts. These asymmetries recommend question-specific calibration: for movements where layout conventions are rigid but color usage is idiosyncratic, we weight palette and semantic channels slightly higher; where chroma is constrained by medium (e.g., cobalt-dominant palettes), we weight layout and semantic channels more strongly. In all cases, the policy discourages "score-only" determinations by requiring a short narrative justification that names the visual regularities actually shared and links those regularities to the channels that fired (Chapter 3.3). [49]

Finally, we adopt a worked-example check for borderline cases: if a candidate pair crosses screening but not referral, the system generates two counter-pairs—(a) a near-miss from the same movement and (b) a palette-matched but layout-divergent control—and displays scores for all three side-by-side. Reviewers then ask whether the candidate's advantage survives immediate neighborhood comparison. If the composite edge disappears, the case is downgraded to "monitor"; if it persists, the case proceeds to referral with a succinct, reproducible rationale. This protocol operationalizes the uncertainty handling and resampling discipline described earlier

(Chapter 3.3) and documents the path from initial suspicion to escalation in a way that can be audited after the fact. [37]

5.2. Interpretability, Human Review, and Governance

Because contested authorship turns on reasons that can be inspected and debated, interpretability is the bridge between statistical evidence and accountable judgment. [44, 45] In this project, "interpretability" is defined operationally: explanations must expose the signals the model actually used and make those signals checkable by a reviewer, not merely decorative.

Each flagged case is therefore accompanied by a deliberately redundant evidence packet aligned to the artifacts specified in Chapter 3: (1) channel-level rationales—paired BLIP captions that verbalize salient motifs; saliency overlays indicating which regions drive agreement; and per-metric notes that state whether alignment arose from palette/harmony, texture/style features, or layout/structure; (2) side-by-side visualizations—aligned crops, palette histograms, stroke- or edge-density maps, and curve-skeleton extractions of anchor shapes; and (3) counterfactual probes—palette-scrambled, edge-preserving, or layout-jittered variants that test whether the match is robust to targeted perturbations. [50] These artifacts are presented with the same rationale panel format used in Chapters 3.2–3.3 so that reviewers can verify *where* agreement occurs and *which* channels contribute to it, consistent with the view that interpretability must be actionable rather than ornamental. [44, 45]

The packet is designed to focus the review on answerable questions: If color is perturbed while edges and layout remain, does the match survive? If edges are preserved but layout is shuffled, does it collapse? If captions swap to near-synonyms, do semantic alignments drift? When evidence proves fragile to a targeted probe, the tool recommends "no concern" or "monitor" rather than escalation, and it records that rationale alongside the scores to prevent silent reliance on a single channel (Chapter 3.3). [44]

Decision rights remain human, and the system's posture is supportive rather than accusatory. When a referral is recommended, the dossier includes a locked analysis manifest—model and dataset version hashes, prompt-lexicon snapshots, preprocessing operations, random seeds, and hardware/driver info—and an append-only export of originals and intermediates with trusted timestamps and audit-trail metadata (as used for reproducibility in Chapter 3.3). This enables independent replication, preserves chain-of-custody, and allows another team to recover the same outputs within stated uncertainty bounds. [37] Institutions should codify (i) roles and panels for review, (ii) rebuttal and right-of-response windows for artists, (iii) minimum and maximum retention periods for data and artifacts, and (iv) re-scoring protocols after model upgrades, so that interpretability norms in ML and due-process expectations in cultural and legal forums are jointly satisfied. [45]

To ensure faithfulness—that explanations reflect what the model actually used—we add a lightweight sanity-check suite referenced in Chapter 3: hide-and-seek tests that mask top-saliency regions to confirm score drops; palette-only and edge-only re-scorings to verify channel claims; and text-swap probes using near-synonyms to detect caption drift. Any failure triggers an automatic downgrade to "monitor" and requires curator acknowledgment before further action. This discipline keeps rationales from becoming ornamental and keeps the human-in-the-loop genuinely informative rather than perfunctory. [44, 50]

5.3. Limitations, External Validity, and Next Steps

Three external-validity issues remain central: style-family confounds, domain shift, and the scarcity of authoritative ground truth. The manuscript now states these limits more briefly because the key point is practical: any deployment should recalibrate thresholds against a fresh, locally curated comparison set.

(ii) Domain shift. Semantic encoders trained on web-scale imagery can under-represent specific materials, techniques, or canons; cross-cultural and medium-specific biases may depress or distort similarity. For domains such as sumi-ink, textile, or collage, we re-weight channels toward physically apt descriptors (stroke directionality, weave periodicity, cut-edge topology), curate domain lexicons to stabilize caption quality, and run periodic bias audits with locally labeled sets. Where shifts are substantial, we adopt lightweight domain adaptation and surface a domain-mismatch warning beside the score so that reviewers interpret thresholds appropriately. These practices follow the general insight that distributional differences require explicit treatment for reliable generalization. [51]

Looking forward, two methodological directions are especially promising. First, process-sensitive features—micro-signatures of medium and making (drying-edge halos, pooling/drag artifacts, binder-specific micro-textures)—can help distinguish convergent style from copied procedure, reducing palette overweighting. These features remain within our interpretability norm by logging where in the image they are detected and how they affect the score (Chapter 3.2). Second, causal probes should be routine: if removing a hypothesized motif (or perturbing layout while preserving palette) collapses similarity, claims that the copying is structural—not merely chromatic—gain credibility. [50] Operationally, we recommend: (a) periodic threshold recalibration against freshly curated "near-neighbor but legitimate" controls, so that operating points reflect current collections; (b) ongoing provenance checks to guard against dataset leakage (e.g., a generator trained on the very work under review), with any leakage flag blocking referral pending curator review; and (c) strict evidentiary hygiene—hash-chained logs, trusted timestamps, and reproducible manifests—treated not as bureaucracy but as the precondition for fair challenge and independent replication. [52-54]

6. Conclusion

This study presents a concise, evidence-centered framework for screening plagiarism in GAN-generated abstract art. Its main contribution is not a single new metric, but the combination of perceptual, compositional, and semantic evidence in a form that can be audited by scholars, curators, and legal reviewers.

Empirically, the representative cases show why fusion is necessary: semantic similarity alone can overstate palette echoes, while compositional evidence alone can miss conceptually close cases. The combined index performs better as a practical screening tool, especially when paired with uncertainty intervals, short rationale paragraphs, and image-level explanatory outputs.

The framework should therefore be used as transparent decision support rather than as an automatic infringement detector. Its immediate value lies in disciplined triage, documented review, and provenance-aware comparison; its longer-term value depends on larger, better-cited datasets and continued human oversight.

Abbreviations

GAN Generative Adversarial Network

Author Contributions

Byungkil Choi: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Goodfellow, I. J. Generative adversarial networks. Communications of the ACM. 2020, 63(11), 139-144. https://doi.org/10.1145/3422622

[2] Karras, T., Laine, S., Aila, T. A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, 4401-4410. Available from: https://openaccess.thecvf.com/content_CVPR_2019/html/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.html

[3] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. Generative adversarial nets. arXiv. 2014, arXiv:1406.2661. Available from: https://arxiv.org/abs/1406.2661

[4] Elgammal, A., Liu, B., Elhoseiny, M., Mazzone, M. CAN: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms. Proceedings of the Eighth International Conference on Computational Creativity. 2017, 96-103. Available from: https://arxiv.org/abs/1706.07068

[5] Tate. Abstract art. Available from: https://www.tate.org.uk/art/art-terms/a/abstract-art (accessed 23 October 2025).

[6] The Museum of Modern Art. Abstraction. Available from: https://www.moma.org/collection/terms/abstraction (accessed 23 October 2025).

[7] Tate. Appropriation. Available from: https://www.tate.org.uk/art/art-terms/a/appropriation (accessed 23 October 2025).

[8] Buskirk, M. The contingent object of contemporary art. Cambridge, MA: MIT Press; 2003, pp. 1-18.

[9] Abd Warif, N. B., Wahab, A. W. A., Idris, M. Y. I., Salleh, R., Othman, F. Copy-move forgery detection: Survey, challenges and future directions. Journal of Network and Computer Applications. 2016, 75, 259-278. https://doi.org/10.1016/j.jnca.2016.09.001

[10] Thyagharajan, K. K., Kumar, S. A. P. A review on near duplicate detection of images using computer vision techniques. arXiv. 2020, arXiv:2009.03224. Available from: https://arxiv.org/abs/2009.03224

[11] 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. Proceedings of the 38th International Conference on Machine Learning. 2021, 139, 8748-8763. Available from: https://proceedings.mlr.press/v139/radford21a.html

[12] Li, J., Li, D., Xiong, C., Hoi, S. C. H. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. Proceedings of the 39th International Conference on Machine Learning. 2022, 162, 12888-12900. Available from: https://proceedings.mlr.press/v162/li22n.html

[13] Redies, C., Brachmann, A. Statistical image properties in large subsets of traditional art, bad art, and abstract art. Frontiers in Neuroscience. 2017, 11, 593. https://doi.org/10.3389/fnins.2017.00593

[14] Datta, R., Joshi, D., Li, J., Wang, J. Z. Studying aesthetics in photographic images using a computational approach. In: Leonardis, A., Bischof, H., Pinz, A. eds. Computer Vision - ECCV 2006. Berlin: Springer; 2006, pp. 288-301. https://doi.org/10.1007/11744078_23

[15] U.S. Copyright Office. Works containing material generated by artificial intelligence. Available from: https://www.copyright.gov/ai/ai_policy_guidance.pdf (accessed 23 October 2025).

[16] U.S. Copyright Office. Copyright and artificial intelligence, Part 3: Generative AI training. Available from: https://www.copyright.gov/ai/ (accessed 23 October 2025).

[17] World Intellectual Property Organization. The WIPO conversation on intellectual property and frontier technologies. Available from: https://www.wipo.int/about-ip/en/frontier_technologies/conversation.html (accessed 23 October 2025).

[18] Redies, C. Combining universal beauty and cultural context in a model of aesthetic experience. Frontiers in Human Neuroscience. 2015, 9, 218. https://doi.org/10.3389/fnhum.2015.00218

[19] Gatys, L. A., Ecker, A. S., Bethge, M. Image style transfer using convolutional neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016, 2414-2423. Available from: https://openaccess.thecvf.com/content_cvpr_2016/html/Gatys_Image_Style_Transfer_CVPR_2016_paper.html

[20] Ahmed, I. T., Hammad, B. T., Jamil, N. A comparative analysis of image copy-move forgery detection techniques. Indonesian Journal of Electrical Engineering and Computer Science. 2021, 22(2), 1177-1190. Available from: https://ijeecs.iaescore.com/index.php/IJEECS/article/view/23881

[21] 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. 2004, 13(4), 600-612. https://doi.org/10.1109/TIP.2003.819861

[22] Rubner, Y., Tomasi, C., Guibas, L. J. The Earth Mover's Distance as a metric for image retrieval. International Journal of Computer Vision. 2000, 40(2), 99-121. https://doi.org/10.1023/A:1026543900054

[23] Johnson, C. R. Jr., Hendriks, E., Berezhnoy, I. J., Brevdo, E., Hughes, S. M., Daubechies, I., Li, J., Postma, E., Wang, J. Z. Image processing for artist identification: Computerized analysis of Vincent van Gogh's painting brushstrokes. IEEE Signal Processing Magazine. 2008, 25(4), 37-48. https://doi.org/10.1109/MSP.2008.923513

[24] Redies, C. A universal model of aesthetic perception based on the sensory coding of natural stimuli and art. Vision Research. 2017, 133, 130-144. https://doi.org/10.1016/j.visres.2017.01.004

[25] Hauagge, A. C. B., Snavely, N. Image matching using local symmetry features. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2012, 206-213. https://doi.org/10.1109/CVPR.2012.6247692

[26] Cohen-Or, D., Sorkine, O., Gal, R., Leyvand, T., Xu, Y.-Q. Color harmonization. ACM Transactions on Graphics. 2006, 25(3), 624-630. https://doi.org/10.1145/1141911.1141933

[27] Canny, J. A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1986, 8(6), 679-698. https://doi.org/10.1109/TPAMI.1986.4767851

[28] Hou, X., Zhang, L. Saliency detection: A spectral residual approach. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2007, 1-8. https://doi.org/10.1109/CVPR.2007.383267

[29] Wang, Z., Simoncelli, E. P., Bovik, A. C. Multiscale structural similarity for image quality assessment. Proceedings of the 37th Asilomar Conference on Signals, Systems and Computers. 2003, 1398-1402. https://doi.org/10.1109/ACSSC.2003.1292216

[30] Lin, J. Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory. 1991, 37(1), 145-151. https://doi.org/10.1109/18.61115

[31] Zhang, T. Y., Suen, C. Y. A fast parallel algorithm for thinning digital patterns. Communications of the ACM. 1984, 27(3), 236-239. https://doi.org/10.1145/357994.358023

[32] Huttenlocher, D. P., Klanderman, G. A., Rucklidge, W. J. Comparing images using the Hausdorff distance. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1993, 15(9), 850-863. https://doi.org/10.1109/34.232073

[33] Bigun, J., Granlund, G. H. Optimal orientation detection of linear symmetry. Computer Vision, Graphics, and Image Processing. 1987, 37(1), 23-33. https://doi.org/10.1016/0734-189X(87)90116-4

[34] Field, D. J. Relations between the statistics of natural images and the response properties of cortical cells. Journal of the Optical Society of America A. 1987, 4(12), 2379-2394. https://doi.org/10.1364/JOSAA.4.002379

[35] Itten, J. The elements of color. New York, NY: Van Nostrand Reinhold; 1970, pp. 20-22.

[36] Hastie, T., Tibshirani, R., Friedman, J. The elements of statistical learning: Data mining, inference, and prediction. 2nd ed. New York, NY: Springer; 2009, pp. 119-127.

[37] Efron, B., Tibshirani, R. J. An introduction to the bootstrap. New York, NY: Chapman and Hall/CRC; 1993, pp. 160-186.

[38] WikiArt. Visual art encyclopedia. Available from: https://www.wikiart.org/ (accessed 23 October 2025).

[39] Gonzalez, R. C., Woods, R. E. Digital image processing. 4th ed. Hoboken, NJ: Pearson; 2018, pp. 857-861.

[40] Zadrozny, B., Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2002, 694-699. https://doi.org/10.1145/775047.775151

[41] Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960, 20(1), 37-46. https://doi.org/10.1177/001316446002000104

[42] Fleiss, J. L. Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971, 76(5), 378-382. https://doi.org/10.1037/h0031619

[43] Efron, B. Bootstrap methods: Another look at the jackknife. The Annals of Statistics. 1979, 7(1), 1-26. https://doi.org/10.1214/aos/1176344552

[44] Lipton, Z. C. The mythos of model interpretability. Communications of the ACM. 2018, 61(10), 36-43. https://doi.org/10.1145/3233231

[45] Doshi-Velez, F., Kim, B. Towards a rigorous science of interpretable machine learning. arXiv. 2017, arXiv:1702.08608. Available from: https://arxiv.org/abs/1702.08608

[46] Kendall, A., Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision? arXiv. 2017, arXiv:1703.04977. Available from: https://arxiv.org/abs/1703.04977

[47] Davis, J., Goadrich, M. The relationship between precision-recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning. 2006, 233-240. https://doi.org/10.1145/1143844.1143874

[48] Goodfellow, I., Bengio, Y., Courville, A. Deep learning. Cambridge, MA: MIT Press; 2016, pp. 505-531.

[49] Fawcett, T. An introduction to ROC analysis. Pattern Recognition Letters. 2006, 27(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

[50] Wachter, S., Mittelstadt, B., Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law and Technology. 2018, 31(2), 841-887. Available from: https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf

[51] Ben-David, S., Blitzer, J., Crammer, K., Pereira, F. A theory of learning from different domains. Machine Learning. 2010, 79(1-2), 151-175. https://doi.org/10.1007/s10994-009-5152-4

[52] International Organization for Standardization. ISO/IEC 27037:2012 Information technology - Security techniques - Guidelines for identification, collection, acquisition and preservation of digital evidence. Available from: https://www.iso.org/standard/44381.html (accessed 23 October 2025).

[53] National Institute of Standards and Technology. Special Publication 800-92: Guide to computer security log management. Available from: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-92.pdf (accessed 23 October 2025).

[54] Internet Engineering Task Force. RFC 3161: Internet X.509 Public Key Infrastructure time-stamp protocol (TSP). Available from: https://www.rfc-editor.org/rfc/rfc3161 (accessed 23 October 2025).