Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Detecting Cross-Medium Stylistic Signals in Sketches and Paintings with Pretrained Visual Encoders

Hassan Ugail
h.ugail@bradford.ac.uk
University of Bradford
Jan Ritch-Frel
Independent Media Institute
Irina Matuzava
Independent Media Institute
Christopher Brooke
University of Nottingham

Research Article

Keywords: artist attribution, cross-medium verification, media computing, pretrained visual representations, computer vision, computational art analysis, trustworthy visual AI

Posted Date: June 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10066839/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Detecting Cross-Medium Stylistic Signals in Sketches and Paintings with Pretrained Visual Encoders

Hassan Ugail[1*], Jan Ritch-Frel[2], Irina Matuzava[2], Christopher Brooke[3]

[1*]Centre for Visual Computing and Intelligent Systems, University of Bradford, United Kingdom.
[2*]Independent Media Institute, United States.
[3]Department of History, University of Nottingham, United Kingdom.

*Corresponding author(s). E-mail(s): h.ugail@bradford.ac.uk;

Abstract

We ask whether aspects of an artist's visual style persist across sketches and finished paintings, and how computational methods can be used to investigate this question as a problem of cross-medium visual media understanding without reducing style to a simple problem of attribution. The study is situated within computational art history and media computing, and treats cross-medium authorship analysis as a way of probing the relation between artistic identity, medium, visual form, and perceptual representation. We examine a controlled corpus of 666 works by ten historical artists and test whether computational models can identify when a sketch and a painting are likely to be by the same artist. The evaluation is designed to measure whether the models generalise to artists not seen during training. We compare three pretrained visual encoders, CLIP, DINOv2, and SigLIP, with handcrafted descriptions of composition, tonal organisation, and surface texture. CLIP and SigLIP provide the strongest evidence of cross-medium transfer, with mean AUC values of 0.766 and 0.759, respectively. DINOv2 performs more weakly but remains above chance, with a mean AUC of 0.612. The handcrafted descriptors do not exceed chance under the same conditions. These findings suggest that pretrained visual encoders can capture stylistic signals that persist across sketches and paintings, and can support perception-aware analysis of complex cultural images across media. At the same time, performance drops sharply when the models are tested against visually similar negative examples, which shows that the detected signal is fragile and should not be interpreted as

evidence of reliable authentication. We therefore argue that cross-medium verification is valuable not as a standalone tool for assigning authorship, but as a method for examining how artificial intelligence models visual style and cross-medium similarity. It contributes to debates on the interpretability, reliability, and evaluation of visual representations for complex media.

Keywords: artist attribution, cross-medium verification, media computing, pretrained visual representations, computer vision, computational art analysis, trustworthy visual AI

1 Introduction

Questions of artistic authorship have long occupied art history, museum provenance research, and the study of cultural heritage. Attribution is never only a matter of naming a maker. It also concerns how visual form, technique, material practice, historical context, and expert judgement are brought together to support claims about artistic identity. Computational methods have increasingly entered this field, first through wavelet, texture, and brushstroke analysis [1, 2], then through sparse coding and other feature-based approaches [3], and more recently through deep representation learning, pretrained visual models, and authentication-oriented learning frameworks [4–7, 11, 12]. Closely related to the present study, one-class learning on compact feature representations has recently been used to verify historical sketches, underlining the value of drawings as a distinct object of computational authentication [13]. These developments have shown that computational systems can identify regularities in artworks that are difficult to describe manually. They have also raised questions about what such systems are actually learning, how their outputs should be interpreted, and where their evidential limits lie. From a media computing perspective, artworks are visually rich media objects in which semantic content, perceptual organisation, material appearance, and historical context are closely entangled. This makes them a useful test case for evaluating whether pretrained visual representations can support perception-aware media analysis beyond standard photographic domains.

Here we examine these questions through the problem of cross-medium artistic style. We ask whether aspects of an artist's visual identity persist across sketches and finished paintings, and whether pretrained visual encoders can detect such persistence. The question is relevant to computational art history because sketches and paintings occupy different positions in artistic practice. A sketch may record exploration, preparation, compositional planning, or rapid observation, while a painting usually involves different materials, temporal rhythms, and surface effects. If a computational model identifies a relation between the two, the result cannot be explained simply as a match of colour, texture, or medium-specific facture. It instead points to a more abstract form of visual similarity that may involve composition, spatial organisation, handling of form, or other recurrent properties of artistic practice. In this sense, the task also speaks to a broader question in visual media analysis: whether learned representations preserve semantically and perceptually meaningful structure across changes in medium.

The cross-medium setting has practical significance for provenance and heritage research, and it also provides a demanding media-analysis setting, since preparatory drawings often survive separately from the works to which they may be historically related. At the same time, it presents a substantial methodological challenge. Most computational attribution studies remain within a single medium or a relatively homogeneous image domain. In such settings, models may benefit from medium-specific cues, digitisation artefacts, colour palettes, or surface texture. A sketch-to-painting comparison is more demanding because it requires a representation that can connect works across a large visual and material gap. This challenge is related to broader work on domain adaptation and cross-domain retrieval, where the central question is whether class-relevant structure survives a shift in visual domain [21–23]. In the present study, however, the aim is not to learn an explicit domain alignment. It is to test whether artist-relevant structure is already present in frozen pretrained visual representations.

To investigate this question, we use a controlled corpus of 666 works by ten historical artists. We formulate the task as cross-medium pairwise verification. Given a sketch and a painting, the model estimates whether the two works are by the same artist. This formulation is appropriate because the problem is relational rather than classificatory. It tests whether a representation preserves artist-specific similarity across media, rather than whether it can assign an image to one of a fixed set of known classes. The resulting protocol therefore provides an evaluation setting for cross-medium visual representations under ordinary, period-controlled, and visually confounded comparisons. We evaluate generalisation using a leave-one-artist-out protocol, in which models are trained on nine artists and tested on a held-out tenth artist. This design is intended to reduce the risk that performance reflects memorisation of the artists in the training set.

The study compares three pretrained visual encoders, CLIP ViT-B/32 [8], DINOv2 ViT-S/14 [9], and SigLIP ViT-B/16 [10], with three handcrafted descriptor families designed to capture composition and form, tonal organisation, and surface texture. The comparison is intended to address both performance and interpretation. If only a single pretrained model produced positive results, the effect might be tied to that model's architecture, training data, or objective. If several distinct encoders produced similar results, the evidence for a reproducible cross-medium signal would be stronger. The inclusion of DINOv2 is especially important because it provides a vision-only comparison point that does not depend on language supervision. The handcrafted descriptors serve a different purpose. They are not treated as state-of-the-art competitors, but as interpretable probes for possible visual cues. The comparison also recognises that stronger perceptual and structure-aware descriptors remain important future baselines [25, 26]. This allows the study to connect performance with questions of interpretability, robustness, and the evaluation of visual media representations.

In this work, we make three contributions. First, it provides a systematic empirical study of cross-medium artist verification between sketches and paintings. Second, it shows that pretrained visual encoders can detect a cross-medium stylistic signal that generalises to artists not seen during training. In the leave-one-artist-out evaluation, CLIP and SigLIP achieve mean AUC values of 0.766 and 0.759 respectively, while DINOv2 achieves a lower but still above-chance mean AUC of 0.612. The close

agreement between CLIP and SigLIP is especially notable because the models differ in training objective and architecture. Third, in this paper, we identify a clear limit to the method. Performance drops sharply when the models are tested against visually similar negative examples. This result shows that the detected signal is fragile and should not be interpreted as evidence of reliable authentication.

Our argument here is therefore deliberately cautious. Cross-medium verification is valuable not as a standalone tool for assigning authorship, but as a way of examining how artificial intelligence models visual style. The results suggest that pretrained encoders capture aspects of artistic similarity that are not recovered by the handcrafted descriptors tested here. They also show that such similarity is unstable under more difficult visual comparisons. This combination of positive evidence and empirical limitation is central to the contribution of the study. Rather than treating computational performance as a substitute for art-historical judgement, we use performance, failure, and model comparison to clarify what kinds of stylistic relation current visual encoders can and cannot represent.

The remainder of the paper is organised as follows. The Materials and Methods section describes the corpus, feature representations, pair construction, evaluation protocol, and statistical analysis. The Results and Discussion section reports the cross-medium verification results and interprets them in relation to computational art history, model comparison, and the limits of attribution-oriented inference. The Conclusion summarises the findings and outlines directions for larger, more diverse, and more interpretable studies of computational approaches to art.

2 Materials and Methods

2.1 Corpus

The study uses a controlled corpus of 666 digitised artworks by ten historical artists. For each artist, the corpus contains sketches and paintings, with 333 sketches and 333 paintings in total. The artists span five broad historical groupings and were selected to provide at least twenty works in each medium while covering variation in period, national tradition, and subject matter. Table 1 reports the composition of the corpus.

The corpus is designed to support a cross-medium question rather than a conventional closed-set classification task. Each artwork is represented by its image, its medium, either a sketch or a painting, and its artist identity. The central task is to determine whether a sketch and a painting are likely to have been produced by the same artist. This design reflects the art-historical problem more directly than assigning every image to one of a fixed set of known artists.

Artists belonging to a period represented by only one individual, namely Canaletto and Ingres, are excluded from period-matched negative sampling but are retained in all other evaluations. This prevents period-matched comparisons from becoming artist-identical by construction.

Table 1: Dataset composition. S = sketches, P = paintings.

Period	Artist	S	P	Total
Renaissance	Leonardo da Vinci (1452–1519)	42	24	66
	Michelangelo (1475–1564)	30	31	61
	Raphael Sanzio (1483–1520)	31	30	61
	Peter Paul Rubens (1577–1640)	30	34	64
Baroque[1]	G. A. Canal / Canaletto (1697–1768)	33	30	63
Neoclassical[1]	J.-A.-D. Ingres (1780–1867)	32	31	63
Impressionist	Georges Seurat (1859–1891)	39	45	84
	John Singer Sargent (1856–1925)	34	31	65
Pre-Raphaelite	Edward Burne-Jones (1833–1898)	32	40	72
	Frederic Leighton (1830–1896)	30	37	67
Total		333	333	666

[1]Period with a single artist. Excluded from period-matched negative sampling.

2.2 Cross-medium verification task

Let $D = \{(x_i, m_i, y_i, p_i)\}_{i=1}^{N}$ denote the corpus. The vector $x_i \in \mathbb{R}^{d_\phi}$ is the feature representation produced by encoder ϕ, the variable $m_i \in \{S, P\}$ indicates whether the work is a sketch or a painting, the label $y_i \in \{1, \ldots, C\}$ denotes artist identity, and p_i denotes historical period. In this study $C = 10$. All feature vectors are ℓ_2 normalised, and cosine similarity is defined as,

$$\kappa(u, v) = u^\top v. \tag{1}$$

For artist c and encoder ϕ, the cross-medium similarity gap is defined as,

$$\hat{\Delta}_\phi(c) = \bar{\kappa}_W^\phi(c) - \bar{\kappa}_B^\phi(c), \tag{2}$$

where $\bar{\kappa}_W^\phi(c)$ is the mean cosine similarity over same-artist sketch and painting pairs, and $\bar{\kappa}_B^\phi(c)$ is the mean cosine similarity over pairs in which the sketch is by artist c and the painting is by another artist.

The verification task asks whether a sketch and a painting are by the same artist. A model is given a sketch and painting pair and returns a probability of shared authorship. The primary evaluation uses a leave-one-artist-out protocol. In each fold, the verifier is trained on nine artists and tested on the held-out artist. This protocol is stricter than closed-set artist classification because it tests whether the learned relation transfers to an artist identity not present during training.

Three negative pair types are used. Unmatched negatives pair a sketch with a randomly selected painting by another artist. Period-matched negatives restrict the negative painting to an artist from the same historical period. Hard negatives are selected by nearest-neighbour search and pair each sketch with the most visually similar painting by a different artist in the relevant feature space. The hard negative condition is intended as a stress test rather than as the primary estimate of ordinary performance. This design is related to cross-domain representation learning, but differs from adversarial or alignment-based approaches because the encoders are not trained

to suppress medium-specific information [21, 22]. It is also related to cross-domain retrieval, where generalisation across both domain and identity is central [23].

2.3 Feature representations

Six feature families are evaluated. Three are pretrained visual encoders and three are handcrafted descriptor families. All pretrained encoder weights remain frozen throughout the experiments.

CLIP ViT-B/32 [8] produces a 512-dimensional visual embedding from a vision and language model trained on large-scale image and text pairs. Only the image encoder is used in this study.

DINOv2 ViT-S/14 [9] produces a 384-dimensional embedding from a vision-only self-supervised model. It is included because it allows us to test whether cross-medium performance depends on language supervision or whether the effect is also present in a purely visual representation.

SigLIP ViT-B/16 [10] produces a 768-dimensional embedding using a sigmoid contrastive objective. It differs from CLIP in both loss function and model configuration, which makes it useful for testing whether any observed result is specific to CLIP or is reproducible across different pretrained encoders.

Composition and Form is a 1398-dimensional descriptor family that combines spatial pyramids of edge energy and luminance, HOG features, contour curvature histograms, low-frequency DCT coefficients, and symmetry and balance features. It is intended to test whether compositional habits and spatial organisation can support cross-medium verification.

Tonal Organisation is a 69-dimensional descriptor family that combines luminance quantiles, multiscale pyramid means, shadow and highlight occupancy, radial DCT energy bands, and local tonal standard deviation histograms. It is intended to test whether characteristic distributions of light and dark persist across sketches and paintings.

Surface Texture is a 92-dimensional descriptor family that combines LBP histograms, Gabor filter responses, wavelet sub-band energies, and GLCM statistics. It is expected to be the least transferable family because the material surfaces of drawings and paintings differ substantially.

The handcrafted families are included as interpretable probes rather than as equal-capacity competitors to the pretrained encoders. Their role is to test whether particular art-historical cue hypotheses, such as composition, tone, or surface texture, can independently explain the cross-medium signal. Stronger learned perceptual descriptors and image-structure descriptors remain important directions for future work [25, 26].

2.4 Pairwise verifier

For each feature family, a pairwise verifier is trained to estimate the probability that a sketch and painting are by the same artist. Given a sketch feature vector x_s and a

painting feature vector x_p, the pair descriptor is,

$$d(x_s, x_p) = \big(\kappa(x_s, x_p),\ \|x_s - x_p\|,\ |x_s - x_p|,\ x_s \odot x_p\big). \tag{3}$$

This descriptor combines cosine similarity, Euclidean distance, elementwise absolute difference, and elementwise product. These terms capture complementary aspects of similarity, including angular alignment, overall distance, dimensionwise discrepancy, and co-activation, and provide a compact way of evaluating how different visual media representations encode cross-medium resemblance. A logistic regression classifier with balanced class weighting is trained on these descriptors. Principal component analysis is applied before classification, reducing the descriptor to $\min(64, d-1)$ dimensions. The verifier is identical across feature families so that differences in performance can be attributed to the representations rather than to the classifier.

2.5 Metric learning as a closed-set upper bound

As a secondary analysis, a triplet projection head is trained on CLIP features. This analysis is not part of the leave-one-artist-out generalisation test. It instead measures how well the CLIP representation can be rectified when all artist identities are known during training. The projection function is,

$$f_\theta(x) = \frac{g_\theta(x)}{\|g_\theta(x)\|}, \qquad g_\theta : 512 \to 256 \to 128 \to 128. \tag{4}$$

The projection network uses batch normalisation, ReLU activations, and dropout with probability 0.2. It is trained with triplet loss, using sketches as anchors, same-artist paintings as positives, and different-artist paintings as negatives. Training uses 4096 triplets per epoch for 200 epochs, a margin of $\alpha = 0.5$, and AdamW optimisation with an initial learning rate of $\eta_0 = 10^{-3}$ and cosine annealing. Since all ten artist identities are used during training, this result is interpreted only as a closed-set ceiling.

This analysis is informed by work on metric learning and representation evaluation, including large-margin neighbourhood learning, lifted structured losses, proxy-based objectives, multi-similarity learning, supervised contrastive learning, and controlled evaluation practices [14–20]. It is included to distinguish what is possible in a fixed roster setting from what transfers to unseen artists.

2.6 Statistical analysis

AUC values are reported with 95 per cent bootstrap confidence intervals computed from 2000 resamples. Permutation p values for pairwise verification are computed from 5000 label shuffles. These permutation values are reported without correction across the six feature families and three negative tiers, and are therefore treated as exploratory.

The formal artist-level test is a one-sided Wilcoxon signed-rank test on the ten per-artist similarity gaps. The mean gap $\bar{\Delta}$ is reported as a descriptive summary. A paired two-sided Wilcoxon test is used to compare the trained verifier with the raw cosine baseline.

Representational similarity analysis is performed using the Mantel matrix correlation test [27]. This test compares artist-level dissimilarity matrices using a permutation procedure. It is used here to examine whether different feature families organise the ten artists into similar relational geometries. This use of representational similarity analysis follows broader methodological practice in the analysis of representational spaces [24, 28].

A final confound regression is conducted at the artist-pair level. It controls for same-period status and for a CLIP-derived subject match score based on broad subject categories. Since only ten same-artist entries are available among ninety artist-pair observations, this regression is treated as a conservative diagnostic rather than as a confirmatory model.

3 Results and Discussion

3.1 Negative controls

Before evaluating cross-medium verification, two controls are used to test the integrity of the pipeline. In the shuffled label control, artist labels are randomly reassigned before pair construction. All six feature families collapse to AUC values near 0.500, with values between 0.494 and 0.502. This indicates that the verification pipeline does not produce above-chance results when artist identity has been removed.

The second control tests medium decodability. A linear classifier is trained to distinguish sketches from paintings using five-fold cross-validation. The pretrained encoders achieve very high accuracy, with 97.9 per cent for CLIP, 97.2 per cent for DINOv2, and 98.0 per cent for SigLIP. The handcrafted families achieve between 82.5 per cent and 91.8 per cent accuracy. These results confirm that the medium gap is strongly encoded in all feature spaces. Cross-medium artist verification therefore cannot be explained by a collapse of sketches and paintings into a shared image domain.

3.2 Pairwise verification

The pairwise verification results show that pretrained visual encoders detect a cross-medium same-artist signal under ordinary negative sampling. Under unmatched negatives, CLIP achieves an AUC of 0.703 with a 95 per cent confidence interval of $[0.693, 0.714]$. SigLIP achieves 0.692 with a confidence interval of $[0.682, 0.702]$. DINOv2 achieves 0.601 with a confidence interval of $[0.589, 0.613]$. All three pretrained encoders are above chance under this condition.

Under period-matched negatives, the pretrained encoders maintain above-chance performance. CLIP remains strongest with an AUC of 0.680, followed by SigLIP at 0.635 and DINOv2 at 0.594. The persistence of the effect under period-matched sampling is important because it reduces the possibility that the result is driven only by broad historical period differences.

The handcrafted families do not show comparable performance. Tonal Organisation reaches an AUC of 0.530 under unmatched negatives, but the effect is small and should not be interpreted as a robust artist-level result. Composition and Surface Texture remain at or near chance. This pattern suggests that the pretrained encoders capture

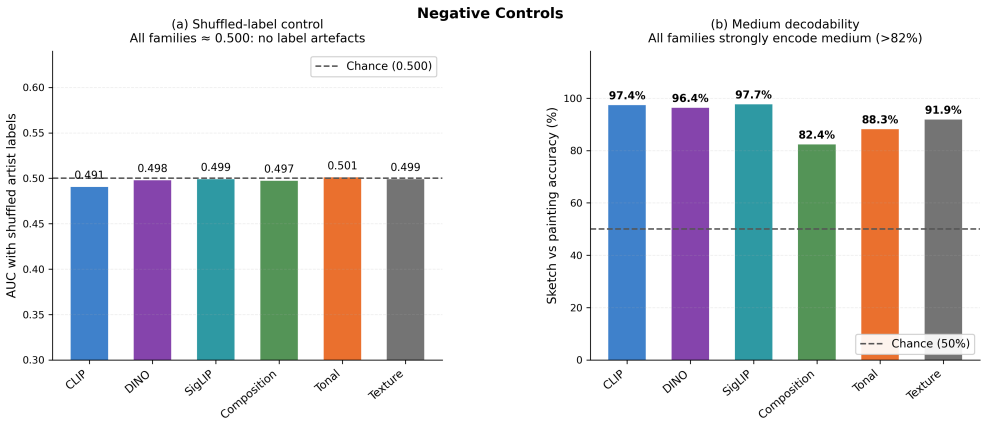


Fig. 1: Negative controls. Panel (a) shows the shuffled label control in which all six feature families collapse to AUC near 0.500. Panel (b) shows medium decodability in which all families achieve more than 82 per cent accuracy at distinguishing sketches from paintings.

Table 2: Pairwise verification AUC and Cohen's d.

Family	Unmatched AUC	Unmatched d	Period AUC	Period d	Hard AUC	Hard d
CLIP	0.703[1]	+0.74	0.680[1]	+0.65	0.469	−0.10
DINOv2	0.601[1]	+0.37	0.594[1]	+0.33	0.372	−0.45
SigLIP	0.692[1]	+0.68	0.635[1]	+0.46	0.518	+0.07
Composition	0.480	−0.05	0.480	−0.07	0.233	−0.61
Tonal	0.530[1]	+0.12	0.523[1]	+0.09	0.466	−0.11
Texture	0.521[1]	+0.01	0.516[1]	+0.00	0.486	−0.05

[1]Permutation $p < 0.05$, uncorrected and exploratory. The formal artist-level test is the one-sided Wilcoxon reported in Table 3.

a form of cross-medium similarity that is not recovered by the handcrafted descriptors tested here.

The hard negative condition establishes the clearest empirical boundary. When each sketch is paired with a visually similar painting by another artist, all feature families collapse to chance or below. CLIP falls to 0.469, DINOv2 to 0.372, and SigLIP to 0.518. The handcrafted families also fail under this condition. This does not invalidate the positive results under unmatched and period-matched negatives. It shows that the signal is statistical and population-level rather than robust for every visually confounded pair. For art-historical use, this boundary is crucial. The models detect a transferable stylistic signal, but they do not provide authentication-level evidence.

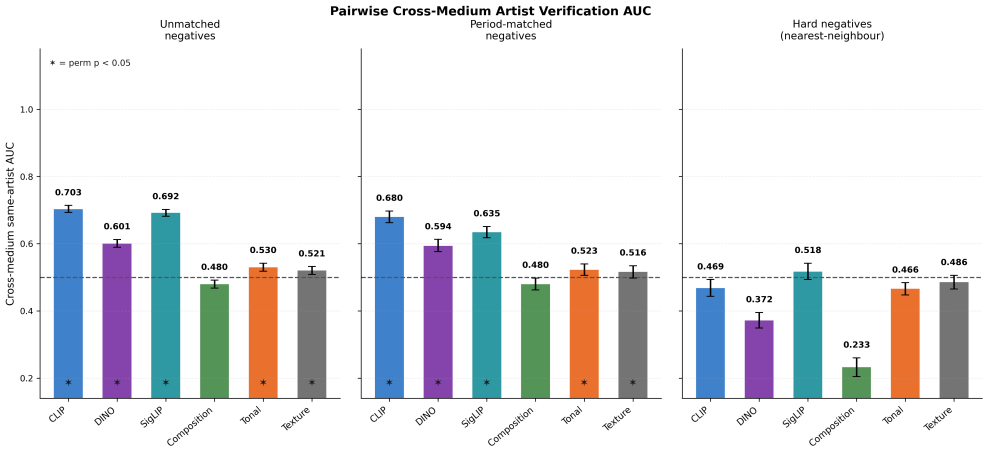


Fig. 2: Pairwise cross-medium same-artist verification AUC across the three negative tiers. Error bars show 95 per cent bootstrap confidence intervals. The asterisk symbol marks permutation $p < 0.05$ without correction. The pretrained families are above chance under unmatched and period-matched negatives. All families collapse under hard negatives.

3.3 Artist-level effects

The artist-level analysis supports the pairwise results. The similarity gap is positive and statistically significant for all three pretrained encoders. CLIP has a descriptive mean gap of $\bar{\Delta} = +0.052$ with a Wilcoxon p value of 0.001 and positive gaps for all ten artists. DINOv2 has a mean gap of $\bar{\Delta} = +0.054$ with $p = 0.003$ and positive gaps for nine of ten artists. SigLIP has a mean gap of $\bar{\Delta} = +0.056$ with $p = 0.002$ and positive gaps for nine of ten artists.

None of the handcrafted families reaches significance at the artist level. Composition has a mean gap close to zero, Tonal Organisation shows a weak positive trend, and Surface Texture is effectively zero. This result is informative because the handcrafted families were designed around plausible art-historical cue hypotheses. Their failure suggests that the transferable signal is not reducible to the particular measurements of composition, tone, and texture used here.

3.4 Leave-one-artist-out generalisation

The leave-one-artist-out evaluation is the primary test of generalisation. CLIP achieves a mean AUC of 0.766 with a 95 per cent confidence interval of $[0.723, 0.818]$ and a one-sided Wilcoxon p value of 0.001. SigLIP achieves a mean AUC of 0.759 with $p = 0.001$. The similarity between these two values is one of the strongest findings in the study. CLIP and SigLIP differ in training objective, patch size, and training data, yet they produce nearly identical transfer performance. This suggests that the detected signal is not an artefact of one specific model family.

Table 3: Artist-level per-gap summary. $\bar{\Delta}$ is the arithmetic mean of ten per-artist gaps and is descriptive only. p is the one-sided Wilcoxon signed-rank test. n_+ counts artists with a positive gap.

Family	$\bar{\Delta}$	se	p_{Wilcoxon}	n_+
CLIP	+0.052	0.011	0.001	10/10
DINOv2	+0.054	0.012	0.003	9/10
SigLIP	+0.056	0.014	0.002	9/10
Composition	−0.002	0.005	0.652	5/10
Tonal	+0.026	0.015	0.080	7/10
Texture	−0.000	0.000	0.348	7/10

Table 4: LOAO verification results. p is the one-sided Wilcoxon p value against chance. Paired p tests whether the trained verifier improves over the raw cosine baseline.

Family	Mean AUC	95% CI	p	Raw AUC Δ	Paired p
CLIP	0.766	[0.723, 0.818]	0.001	0.702 +0.064	0.006
DINOv2	0.612	[0.555, 0.680]	0.003	0.607 +0.005	0.922
SigLIP	0.759	[0.695, 0.830]	0.001	0.705 +0.054	0.006
Composition	0.508	[0.450, 0.560]	0.500	0.473 +0.035	0.557
Tonal	0.530	[0.490, 0.572]	0.161	0.534 −0.004	0.695
Texture	0.537	[0.488, 0.580]	0.116	0.515 +0.022	0.275

DINOv2 achieves a weaker but still above-chance mean AUC of 0.612 with $p = 0.003$. Since DINOv2 is trained without language supervision, its performance indicates that the cross-medium signal cannot be attributed solely to text and image alignment. The result supports the interpretation that pretrained visual systems can encode artist-related structure that extends beyond medium-specific visual appearance.

The trained verifier improves over the raw cosine baseline for CLIP and SigLIP, with paired Wilcoxon p values of 0.006 in both cases. This indicates that the richer pair descriptor contributes useful information beyond direct cosine similarity. For DINOv2, the trained verifier does not improve over the raw cosine baseline. This suggests that the relevant signal in DINOv2 may lie primarily in the directional structure of the embedding itself.

The handcrafted descriptors do not transfer above chance in the leave-one-artist-out setting. Composition achieves 0.508, Tonal Organisation achieves 0.530, and Surface Texture achieves 0.537. These results reinforce the conclusion that the tested handcrafted cues do not independently account for the cross-medium artist signal.

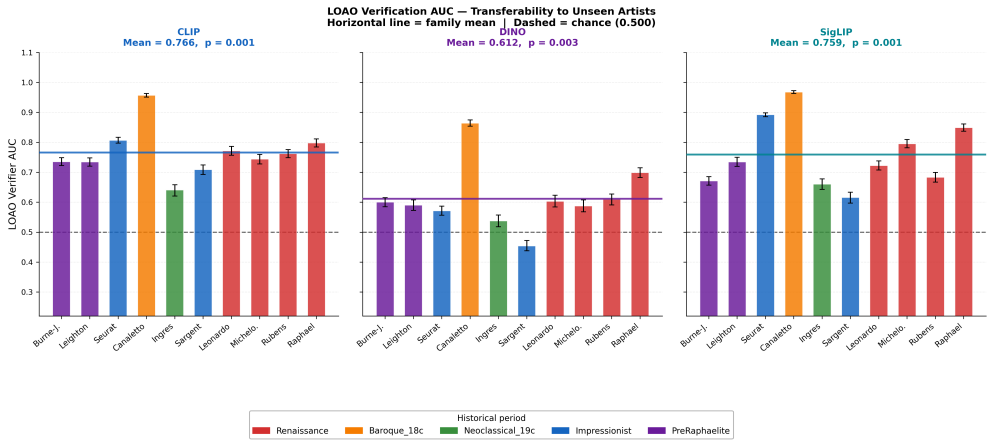


Fig. 3: Per-artist LOAO AUC for the pretrained families CLIP, DINOv2, and SigLIP. Bar colour indicates historical period. The horizontal line marks the family mean and the dashed line marks chance at 0.500. CLIP and SigLIP show consistently strong above-chance performance across artists, with Canaletto exceeding an AUC of 0.96 for CLIP and 0.97 for SigLIP.

3.5 Representational similarity

Representational similarity analysis shows that the three pretrained encoders organise artists into related cross-medium similarity geometries. CLIP and DINOv2 have a Mantel correlation of $r = 0.658$ with $p < 0.001$. CLIP and SigLIP have a correlation of $r = 0.615$ with $p < 0.001$. DINOv2 and SigLIP have a correlation of $r = 0.459$ with $p = 0.005$.

These correlations suggest that the pretrained encoders do not merely achieve similar verification scores. They also structure the relationships among artists in partly similar ways. This strengthens the interpretation that the encoders capture a reproducible relational signal in the corpus. In this sense, the analysis follows the broader logic of representational similarity analysis, where comparison of relational geometries can help interpret what different systems encode [24, 28].

The correlations between pretrained and handcrafted families are much weaker. SigLIP shows limited alignment with Composition and Tonal Organisation, but these correlations do not imply that the handcrafted features recover the verification signal. They indicate only partial similarity in artist-level relational geometry. Overall, the representational similarity analysis supports the distinction between pretrained embeddings and the handcrafted cue families.

3.6 Closed-set metric rectification

The secondary triplet projection analysis shows that CLIP features contain enough information for perfect within-corpus discrimination when all artist identities are known during training. The projected representation achieves top-one retrieval accuracy of 100 per cent and a Fisher criterion gain of approximately 8.5×.

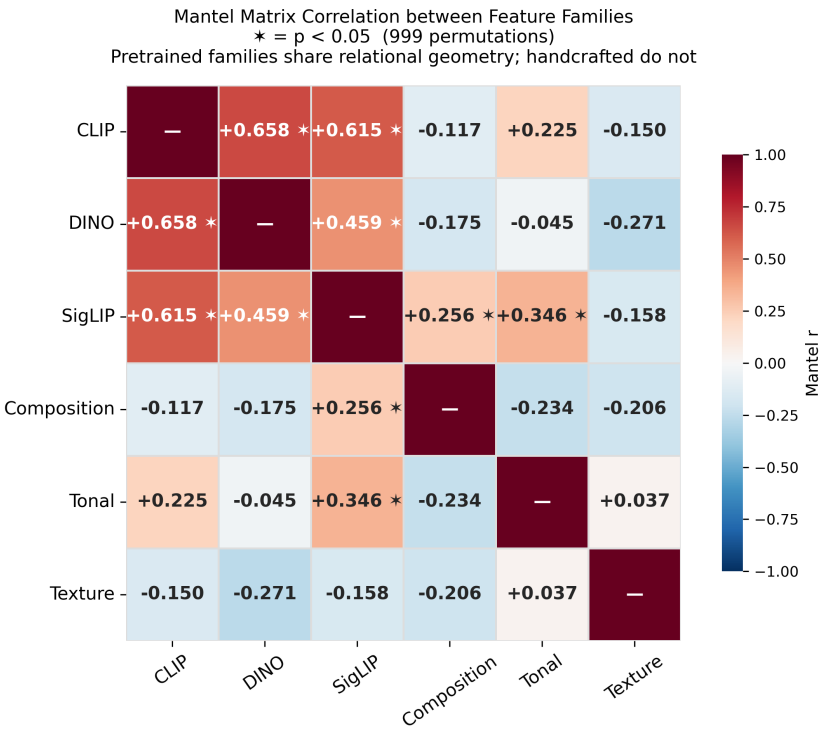


Fig. 4: Mantel matrix correlation coefficients between the feature family artist-level cross-medium similarity matrices. The asterisk symbol marks $p < 0.05$ based on 999 permutations. The pretrained families share consistent relational geometry with correlations between 0.44 and 0.64. No pretrained family correlates significantly with Composition or Texture.

This result should not be confused with the leave-one-artist-out findings. It is a closed-set result and does not show generalisation to unseen artists. Its value is diagnostic. It demonstrates that, within a fixed roster, CLIP features can be reshaped into a highly discriminative cross-medium space. The more important result for art-historical interpretation remains the leave-one-artist-out evaluation, where the model must transfer to an artist not seen during training.

3.7 Implications for computational art history and visual media analysis

The results support a cautious claim. Pretrained visual encoders can detect artist-specific stylistic signals that persist across sketches and paintings, but those signals are fragile and should not be treated as direct evidence of authorship. This distinction is central to the contribution of the study. The models are not reliable authentication systems. They are instruments for probing how visual style is represented computationally and for evaluating how visual media representations behave under controlled

Table 5: Secondary closed-set result comparing raw and triplet projected CLIP using kNN-1 retrieval from sketch to painting. All metrics reflect within-corpus performance only.

Space	Accuracy	MCC	Fisher J	NP@1
Raw CLIP (512 D)	42.9 %	0.372	0.402	—
Projected (128 D)	100.0 %	1.000	3.436	1.000
Fisher gain	—	—	8.5×	

cross-medium shifts. This position is consistent with recent work on computational authentication, which shows the promise of machine learning while also requiring careful attention to evidential limits [11, 12].

The cross-medium setting is especially useful because it reduces dependence on medium-specific cues and places emphasis on more general perceptual and semantic structure. A model that links sketches to paintings cannot rely only on pigment, canvas texture, or other surface properties of finished paintings. It must instead use some more general structure in the image. The present study does not fully explain what that structure is. It may involve composition, spatial organisation, recurring treatment of form, subject conventions, or latent correlations introduced by the corpus. Further work using interpretability methods, intermediate representation probing, and more diverse datasets will be needed to identify the relevant mechanisms.

The failure of the handcrafted families is also important. It does not show that composition, tone, or texture are irrelevant to artistic style. It shows that the specific computational operationalisations used here do not capture the transferable signal. This is an important negative result for computational humanities and media computing because it demonstrates the gap between domain concepts and measurable image descriptors. Terms such as composition and tonal organisation do not become adequate computational variables simply by attaching hand-engineered features to them.

4 Conclusion

In this work, we examine whether artist-specific stylistic signals persist across sketches and finished paintings, and whether pretrained visual encoders can detect such signals as part of a cross-medium visual media analysis task. Using a controlled corpus of 666 works by ten historical artists, in this study, we formulated the problem as cross-medium pairwise verification and evaluated generalisation using a leave-one-artist-out protocol.

The main finding is that pretrained visual encoders detect a transferable cross-medium signal. CLIP and SigLIP achieve mean leave-one-artist-out AUC values of 0.766 and 0.759, while DINOv2 achieves a lower but still above-chance value of 0.612. The close agreement between CLIP and SigLIP, together with the above-chance performance of the vision-only DINOv2 model, suggests that the result is not specific to a single architecture or solely dependent on language supervision.

The handcrafted Composition and Form, Tonal Organisation, and Surface Texture descriptors do not achieve above-chance leave-one-artist-out performance. This negative result is informative because it shows that the tested interpretable descriptors do not explain the transferable signal captured by the pretrained encoders. Future work should therefore develop more structure-aware and art-historically meaningful descriptors rather than treating current handcrafted features as adequate proxies for style.

The study also identifies a clear boundary. All methods fail under hard nearest-neighbour negatives. The detected signal is therefore real but fragile. It is present at the population level, but it is not robust enough for authentication-level attribution. Cross-medium verification should accordingly be understood not as a standalone tool for assigning authorship, but as a method for investigating how artificial intelligence models visual style.

The broader contribution of the paper is to show how computational experiments can clarify both the possibilities and limits of computational approaches to art and complex visual media. The results suggest that pretrained visual encoders encode aspects of artistic similarity that persist across medium, while also demonstrating that such encodings may remain unstable, opaque, and dependent on corpus design. Larger and more diverse datasets, stronger control of digitisation and subject matter, and more interpretable modelling will be necessary before cross-medium computational evidence can play a stronger role in art-historical, provenance, and trustworthy media-analysis settings.

Acknowledgments

The author would like to thank the Centre for Visual Computing and Intelligent Systems at the University of Bradford for providing the computing resources used in this work.

Data availability

Code for the associated feature extraction is publicly available at the GitHub repository: https://github.com/ugail/sketch-verification.

Funding

No specific funding was received for this work.

Author information

Authors and Affiliations

University of Bradford, United Kingdom.
Hassan Ugail
Independent Media Institute, United States.
Jan Ritch-Frel and Irina Matuzava

University of Nottingham, United Kingdom
Christopher Brooke

Corresponding author

Correspondence to Hassan Ugail.

Contributions

H.U. led the research conceptualisation and methodology development. J.R. and I.M. curated and verified datasets. All authors reviewed the results, discussed the experimental findings, and contributed to the interpretation. All authors reviewed the manuscript, and approved the final manuscript.

Declarations

Ethical approval

Not applicable.

Consent to participate

Not applicable.

Consent to publish

Not applicable.

Competing interests

The authors declare no competing interests.

References

[1] Lyu, S., Rockmore, D., Farid, H. (2004). A digital technique for art authentication. *Proceedings of the National Academy of Sciences*, 101(49), 17006–17010.

[2] Johnson, C. R., Hendriks, E., Berezhnoy, I. J., Brevdo, E., Hughes, S. M., Daubechies, I., Li, J., Postma, E., Wang, J. Z. (2008). Image processing for artist identification: computerized analysis of Vincent van Gogh's painting brushstrokes. *IEEE Signal Processing Magazine*, 25(4), 37–48.

[3] Hughes, J. M., Graham, D. J., Rockmore, D. N. (2010). Quantification of artistic style through sparse coding analysis in the drawings of Pieter Bruegel the Elder. *Proceedings of the National Academy of Sciences*, 107(4), 1279–1283.

[4] Cetinic, E., Lipic, T., Grgic, S. (2018). Fine-tuning convolutional neural networks for fine art classification. *Expert Systems with Applications*, 114, 107–118.

[5] Elgammal, A., Liu, B., Kim, D., Elhoseiny, M., Mazzone, M. (2018). The shape of art history in the eyes of the machine. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

[6] Conde, M. V., Turgutlu, K. (2021). CLIP-Art: contrastive pre-training for fine-grained art classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 3956–3960.

[7] Ugail, H., Stork, D. G., Edwards, H. G. M., Seward, S. C., Brooke, C. (2023). Deep transfer learning for visual analysis and attribution of paintings by Raphael. *Heritage Science*, 11, 268.

[8] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, 139, 8748–8763.

[9] Oquab, M., Darcet, T., Moutakanni, T., et al. (2023). DINOv2: learning robust visual features without supervision. *arXiv preprint* arXiv:2304.07193.

[10] Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L. (2023). Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11941–11952.

[11] Ugail, H., Edwards, H. G. M., Omar, Z. (2024). A one-class autoencoder for porcelain art authentication. In *Proceedings of the 2024 IEEE 8th International Conference on Signal and Image Processing Applications (ICSIPA)*. IEEE.

[12] Smith, J. H., Holt, C., Smith, N. H., Taylor, R. P. (2024). Using machine learning to distinguish between authentic and imitation Jackson Pollock poured paintings: A tile-driven approach to computer vision. *PLOS ONE*, 19(6), e0302962.

[13] Ugail, H., Ritch-Frel, J., Matuzava, I., Stork, D. G. (2026). Verification of historical sketches via one-class learning on compact feature representations. *PLOS ONE*, 21(6), e0344796.

[14] Weinberger, K. Q., Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10, 207–244.

[15] Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S. (2016). Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4004–4012.

[16] Wang, X., Han, X., Huang, W., Dong, D., Scott, M. R. (2019). Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5022–5030.

[17] Movshovitz-Attias, Y., Toshev, A., Leung, T. K., Ioffe, S., Singh, S. (2017). No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision*, 360–368.

[18] Kim, S., Kim, D., Cho, M., Kwak, S. (2020). Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3238–3247.

[19] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D. (2020). Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33, 18661–18673.

[20] Musgrave, K., Belongie, S., Lim, S.-N. (2020). A metric learning reality check. In *Proceedings of the European Conference on Computer Vision*, 681–699. Springer.

[21] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, F., Laviolette, F., Marchand, M., Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59), 1–35.

[22] Long, M., Cao, Z., Wang, J., Jordan, M. I. (2018). Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, 31. Curran Associates.

[23] Paul, S., Dutta, T., Biswas, S. (2021). Universal cross-domain retrieval: generalizing across classes and domains. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3136–3145.

[24] Kriegeskorte, N., Mur, M., Bandettini, P. A. (2008). Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 4. https://doi.org/10.3389/neuro.06.004.2008.

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

[26] Ding, K., Ma, K., Wang, S., Simoncelli, E. P. (2022). Image quality assessment: unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5), 2567–2581.

[27] Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2), 209–220.

[28] Popal, H., Wang, Y., Olson, I. R. (2019). A guide to representational similarity analysis for social neuroscience. *Social Cognitive and Affective Neuroscience*, 14(11), 1243–1253.