Synthetic Multi-View Learning for Artist Attribution

Loris Bottegoni[1,*,†], Leonardo Cambiotti[1,*,†], Valerio Crocetti[1,2], Yassir Flavio Suarez Sanchez[1], Malvina Halilaj[1] and Aldo Franco Dragoni[1]

[1]*Università Politecnica delle Marche, Via Brecce Bianche, Ancona, 60131, Italy*
[2]*HES-SO Valais-Waillis, 3960 Sierre, Switzerland*

Abstract

Artist attribution represents a complex challenge in fine-grained image classification due to high intra-class variability and stylistic differences among artists from the same period. In this work, we present a framework based on a Synthetic Multi-view Learning Late Fusion approach for the identification of 177 artists selected from the WikiArt dataset. The proposed system integrates nine heterogeneous models, based on both convolutional architectures (ConvNeXt, RegNet) and Vision Transformers (Swin, ViT, MaxViT), each specialized in extracting complementary features. The framework analyzes the artwork through multiple perspectives: chromatic perception in the CIELAB color space, stroke morphology using structural Sobel filters, periodic components in the frequency domain (FFT), and brushstroke details via a local patch-based analysis. The models are combined using a Weighted Late Fusion strategy, in which the weights and Temperature Scaling parameters are optimized via the CMA-ES algorithm. Experimental results demonstrate the effectiveness of the proposed approach, which achieves a Macro F1-Score of 86.73% and an Overall Accuracy of 87.39%. The analysis highlights how integrating structural information and high-resolution local analysis can help resolve stylistic ambiguities, thereby providing a robust system for Computational Art Analysis.

Keywords

Artist Attribution, Multi-view Learning, Late Fusion, Computational Art Analysis, Computer Vision

1. Introduction

The attribution of a painting to its creator (Artist Attribution) using artificial intelligence represents one of the most complex challenges in Computer Vision. Unlike generic object classification, this problem falls under the category of fine-grained classification: the system must not only recognize the semantic content or the artistic movement but also identify the specific artist. This task is inherently difficult due to high intra-class variability where an artist's style evolves through distinct phases and low inter-class variability, as artists from the same school often share visual patterns and compositional structures. This work focuses on the dimension of formalist analysis, often referred to as digital connoisseurship. The underlying premise is that while artist attribution 'in the wild' involves historical provenance and chemical analysis, a formalist computational approach such as the one proposed here provides a quantitative assessment of the artist's style. It focuses on those micro-gestures, stroke morphologies, and chromatic choices that are often subconscious and purely visual. While broader computational frameworks can now incorporate historical metadata and interpretative context, our work isolates these visual features to refine the practice of digital connoisseurship. Consequently, our framework is designed to provide complementary, data-driven evidence to support experts in scholarly debates where traditional historical analysis cannot reach a unanimous consensus. To achieve this, we propose a Synthetic Multi-view Learning Late Fusion approach. Using the WikiArt dataset as a benchmark,

4th Italian Workshop on Artificial Intelligence for Cultural Heritage (IAI4CH 2026, https://ai4ch.di.unito.it/), co-located with the 24th International Conference of the Italian Association for Artificial Intelligence (AIxIA 2026). 6–9 October 2026, Perugia, Italy

*Corresponding author.
†These authors contributed equally.
S1123887@studenti.univpm.it (L. Bottegoni); S1121776@studenti.univpm.it (L. Cambiotti); valerio.crocetti@hevs.ch (V. Crocetti); S1126616@studenti.univpm.it (Y. F. S. Sanchez); m.halilaj@pm.univpm.it (M. Halilaj); a.f.dragoni@staff.univpm.it (A. F. Dragoni)
0009-0000-8144-7098 (L. Bottegoni); 0009-0001-4139-0522 (L. Cambiotti); 0009-0002-3760-6041 (V. Crocetti); 0009-0000-2642-2840 (Y. F. S. Sanchez); 0009-0001-1981-128X (M. Halilaj); 0000-0002-3013-3424 (A. F. Dragoni)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-4304/paper3.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-10-27

the system integrates nine specialized models that analyze the artwork through multiple perspectives (chromatic, structural, and spectral). By capturing these complementary features, the framework aims to reduce uncertainty in the authentication process, while leaving the final decision to the experts. The complete source code is available on GitHub.[1]

2. Related Work

Traditional artwork attribution relied on handcrafted features like color and composition [1]. The introduction of large-scale datasets such as WikiArt [2] shifted the field toward deep learning to capture latent stylistic features [3, 4]. Studies suggest that an artist's identity is defined by non-linear, multi-scale patterns related to brushstroke topography and unintended stylistic elements [5]. To overcome the limitations of global representations, several works such as those by Yang et al. [6] and Ji et al. [5] have proposed approaches based on local analysis, demonstrating that specific image regions contain informative visual cues for artist attribution. In particular, the use of patches has allowed the capture of low-level features such as texture and brushstrokes, which are often attenuated in global models, as also demonstrated by Rodriguez et al. [7]. The work of Strezoski et al. [8] highlighted how multi-task and multi-label approaches improve the robustness and generalization of models on complex artistic domains. Recent advancements in Vision Transformers [9, 10] and modernized CNNs [11] have further enhanced performance in modeling both global and local dependencies. Additionally, self-supervised and multimodal models [12, 13] demonstrate remarkable generalization, learning transferable representations without explicit labels.

To manage these diverse representations, we draw inspiration from the principle of subnet specialization found in Mixture of Experts (MoE) [14]. Although MoE is typically employed to scale model capacity or handle multimodal tasks, its principles inspire strategies for combining heterogeneous models. In this context, Multi-view Learning has emerged as an effective methodology for learning from different representations of the same data, as seen in the work of Yu et al. [15]. Tancik et al. [16] and Rahaman et al. [17], explores the importance of the frequency domain in image analysis. In the artistic context, spectral analysis is relevant for characterizing brushstrokes and repetitive structures. Concurrently, works such as the one by Avanaki et al. [18] show how perceptually uniform representations, such as the CIELAB color space, can be used to model perceptual chromatic variation more faithfully. In this work, we propose a Synthetic Multi-view Learning framework that unifies state-of-the-art models based on heterogeneous architectures with complementary representations derived from local (patch-based), chromatic (CIELAB), and frequency (FFT) analysis.

3. Dataset Methodology

The dataset construction was designed with the dual objective of ensuring a sufficient number of images to train the models and providing a more meaningful representation of diverse artistic styles, structuring the overall process into three main phases: artist selection, artwork clustering per artist, and stratified splitting alongside training set balancing via data augmentation for underrepresented artists. Regarding the artist selection, the complete dataset comprises 177 artists, selected by including only those with more than 70 artworks in the WikiArt dataset who also belong to an art movement represented by at least 5 artists. For the feature extraction and semantic representation, a vector representation (embedding) was extracted for each image using a ResNet-18 network pre-trained on ImageNet, which served as a feature extractor; the final classification layer was removed, and L2 normalization was applied to project the 512-dimensional embeddings onto the unit hypersphere. This representation enables the measurement of semantic similarity between artworks via cosine distance. For the selection of 300 representative images, unsupervised clustering (*K-Means*) with $k = 300$ was applied for every artist with more than 300 artworks, with the objective of selecting a balanced set of images representative of

[1] https://github.com/LeoC-02/Synthetic-Multi-View-Learning-for-Artist-Attribution

the artist's style while avoiding visual redundancies. For each cluster, the image closest to the centroid in the embedding space is selected:

$$x^* = \arg\min_{x_i \in C_j} d_{\cos}(f(x_i), \mu_j) \tag{1}$$

where μ_j denotes the centroid of cluster C_j. To ensure that transient or minority stylistic phases are not excluded, we observe that the embedding space density naturally allocates clusters to any sufficiently distinct stylistic group; by setting $K = 300$, we provide sufficient granularity to capture even secondary stylistic phases, provided they are represented by a non-negligible number of samples in the raw dataset. Subsequently, for the Train/Validation/Test split, the images for each artist are divided into training, validation, and test sets using agglomerative hierarchical clustering based on cosine distance. A distance threshold is preferred over a fixed cluster count to accommodate each artist's intrinsic stylistic diversity. We empirically adopt a threshold of 0.5, verified via qualitative inspection to balance cluster coherence against excessive fragmentation. Each cluster is subsequently divided using an internally stratified 80%, 10%, 10% split across training, validation, and test sets, respectively, promoting the creation of sets that are as representative as possible of the artist's style without near-duplicate leakage. Finally, for the training set balancing and augmentation, a target of 240 images per artist was defined for the training set to ensure class balancing. Under-represented artists are balanced through a controlled augmentation pipeline, which operates exclusively on the training set, including transformations such as rigid rotations (90°, 180°, 270°) and horizontal flipping. All images, both original and augmented, were resized to 448×448 pixels, and a Round-Robin selection algorithm was applied to the results of the hierarchical clustering with a distance threshold of 0.4 to ensure that the augmented images are distributed as equitably as possible across the various sub-styles.

4. Model Architecture

To build the multi-view framework, we selected heterogeneous architectures designed to capture different aspects of the artwork. A standardized training protocol was adopted for all models. The AdamW optimizer was used, selected for its superior regularization capabilities compared to the standard version. Weight decay, employed to mitigate the risk of overfitting, was set to 1×10^{-2} for all configurations. Cross-Entropy Loss was adopted as the loss function. An effective global batch size of 32 samples was fixed. For architectures with higher GPU memory requirements, this value was maintained through the use of Gradient Accumulation. The initial learning rate was set to 5×10^{-5} and managed by a Cosine Annealing learning rate policy. This strategy includes a linear warmup phase (representing 10% of the total steps) to prevent numerical instability during the initial training stages, followed by a cosine decay to ensure smooth convergence toward a local minimum. To counteract the exploding gradient problem, Gradient Clipping was implemented with a maximum threshold of 1.0. Furthermore, an early stopping mechanism with a patience of 5 epochs was adopted. Although Overall Accuracy is a common metric, we chose Macro F1-Score as the primary metric for checkpoint saving. This choice was necessary because the validation set is imbalanced across the various classes. Finally, the models operate on images resized to 224×224, with the exception of Swin, which processes 256×256 images.

4.1. Fine-tuned CLIP ViT-B/16

The first model is based on CLIP ViT-B/16 [13], a transformer model pre-trained via multimodal image-text learning. In this work, we exclusively utilize the visual branch, adapted for the artist classification task through full fine-tuning. CLIP was selected for its ability to extract generalized representations. It is able to capture subtle stylistic nuances and complex artistic signatures

4.2. Six-channel RegNetY-160 with RGB–CIELAB Early Fusion

The second model, focused on analyzing chromatic and perceptual features, employs a RegNetY-160 architecture [19] adapted to operate on a six-channel input (early fusion). This architecture was chosen

for its integration of Squeeze-and-Excitation modules, which enhance the model's ability to weight channel-wise information. Rather than being restricted to the RGB color space, the model receives as input a tensor composed of the concatenation of RGB and CIELAB channels (Fig. 1). To integrate this additional information without losing the knowledge pre-learned on ImageNet, the model's first convolutional layer was modified: the first three channels were initialized with the original weights, while the remaining three were initialized by averaging the RGB channel weights. This approach allows the model to process structural and chromatic information simultaneously from the early stages of the feature extraction process, specializing in the recognition of local chromatic patterns. CIELAB was selected as it models human chromatic perception more uniformly than RGB, effectively separating luminance from color information.

(a) Channel L (b) Channel A (c) Channel B

Figure 1: LAB decomposition applied to the artwork Bacchus.

4.3. Fine-tuned MaxViT-Small

The third model is based on MaxViT-Small [20], a hybrid architecture that combines local convolutions and global self-attention mechanisms. This feature is particularly relevant in the artistic domain, where both local details (e.g., brushstrokes, textures) and global compositional structures coexist. The model is initialized with weights pre-trained on ImageNet and subsequently fine-tuned on the artist classification task. To improve optimization stability, a gradient accumulation strategy and mixed precision training are adopted, allowing for the use of larger effective batches without compromising computational constraints.

4.4. Fine-tuned Swin V2 Base

A further model is based on Swin Transformer V2 [10], a hierarchical architecture that applies self-attention mechanisms within local windows using Shifted Windowing attention. This scheme allows for modeling both local and global dependencies while maintaining limited computational complexity. Furthermore, the model's multi-scale structure enables the extraction of features at different levels of abstraction, making it ideal for the artistic domain by allowing the simultaneous capture of fine details and image composition. The model is initialized with weights pre-trained on the ImageNet-22k dataset and subsequently fine-tuned.

4.5. Swin V2 with Patch-Based Analysis

To explicitly capture the micro-level stylistic signatures often lost in global resizing, we further leverage the Swin Transformer V2 [10] through a specialized patch-based strategy. In this configuration, each image is resized to 512×512 pixels and divided into five fixed 256×256 regions: the four corner quadrants and one overlapping central patch. During the training phase, the model operates at the patch level. For each iteration, we apply a stochastic sampling strategy where one of the five predefined regions (Fig. 2) is selected with a uniform probability of $p = 0.20$. This process acts as a form of regularization, forcing the feature extractor to identify the artist's identity even from partial fragments of the painting. During the inference phase, we adopt an image-level aggregation strategy. For each

image, five patches $\{p_1, \ldots, p_N\}$ (where $N = 5$) are extracted and processed by the model $\mathcal{M}_\theta$. The model outputs a logit vector $z_i = \mathcal{M}_\theta(p_i)$ for each patch. The final prediction $\hat{y}$ is obtained by averaging the logits across all patches and applying the softmax function σ:

$$\hat{y} = \arg\max \left(\sigma \left(\frac{1}{N} \sum_{i=1}^{N} z_i \right) \right) \tag{2}$$

This approach, often referred to as 'Logit Averaging', mitigates the impact of noise from individual patches and provides a more stable estimate of the class distribution compared to averaging posterior probabilities.

(a) Center (b) Bottom left (c) Bottom right (d) Top left (e) Top right

Figure 2: Image patch division

4.6. Fine-tuned DINO ViT-B/16

Unlike supervised models, which tend to focus on discriminative attributes for ImageNet categories, DINO [12] captures structural patterns and more subtle intra-image relationships. This proves effective in the artistic context, where stylistic differences between different classes are often linked to complex texture details and the spatial coherence of the brushwork, rather than solely to semantic content.

4.7. ConvNeXt-based Architectures

ConvNeXt [11] is a convolutional neural network designed to compete with Vision Transformers through the adoption of similar macro-design areas.

4.7.1. Standard Configuration (Baseline)

In its initial version, the network was used in its original three-channel (RGB) architecture to evaluate the model's ability to discriminate between artists based on the chromatic characteristics and standard textures extracted by the pre-trained convolutional filters.

4.7.2. ConvNeXt with Structural Channel

To better capture the stylistic features of the artworks, the model was extended to process a four-channel input. A structural channel, derived from the Sobel gradient magnitude, is appended to the three standard RGB channels (Fig. 3). This channel emphasizes local intensity transitions and fine-grained variations, which are highly indicative of an artist's characteristic brushwork and texture. The structural map focuses exclusively on the artist's handling of the brush by isolating morphology from chromatic interference. The structural map is computed as the Euclidean norm of the horizontal and vertical Sobel derivatives, subsequently normalized and inverted to highlight high-frequency structural details. The model's stem was modified to accommodate this additional input, with the initial weights of the fourth channel calculated by averaging the pre-trained weights of the three RGB channels, ensuring a seamless integration of the structural information. A distinctive feature of the training protocol is the implementation of Channel Dropout: during training, either the RGB stream or the structural channel is randomly omitted with a specific probability. This regularization technique prevents the model

from relying exclusively on one source of information, forcing the network to extract discriminative stylistic features from both the chromatic and structural representations, thereby enhancing the model's generalization capability.

(a) Original image (b) Sobel filter

Figure 3: Comparison between original image and edge detection.

4.7.3. ConvNeXt with FFT

An additional version of the model is represented by the Hybrid ConvNeXt architecture. In this configuration, the network input was extended to six channels to combine traditional visual information with spectral analysis in the frequency domain. Specifically, the three original RGB channels are complemented by three channels containing the frequency spectrum, obtained via the Discrete Fourier Transform (DFT) applied individually to each color band. The spectrum is normalized on a logarithmic scale to stabilize the value distribution and enable the network to identify periodic patterns. The resulting six channels are concatenated and fed as input to the model, whose initial layer is adapted to accept multi-channel input while maintaining an initialization consistent with the pre-trained weights. The spectral analysis identifies periodic patterns and spatial frequencies that characterize both the physical support and the artist's rhythmic application of paint. This is particularly relevant for movements where the repetition of strokes creates a distinct frequency signature.

5. Model Integration and Late Fusion Strategy

To integrate the diverse perspectives captured by the nine models described in Section 4, we have implemented a *Synthetic Multi-view Late Fusion* strategy. The objective is to overcome the limitations of simple arithmetic averaging of predictions by adopting a calibration and weighting approach.

5.1. Logit Extraction and Calibration

The fusion process operates on the raw logits $L \in \mathbb{R}^C$ produced by the final linear layer of each model, where C is the number of classes. Since the architectures utilized exhibit varying levels of confidence, we have introduced a Temperature Scaling phase for each model i.

The calibrated logits L'_i are defined as:

$$L'_i = \frac{L_i}{T_i} \tag{3}$$

where $T_i \in [0.5, 3.0]$ is the temperature parameter that softens or sharpens the probability distribution, aligning the confidence scales of the heterogeneous models. As demonstrated in the work of Guo et al. [21], this technique represents a simple and strong post-hoc calibration method, allowing the logits to be rescaled such that the output probabilities reflect the actual model accuracy without altering the ranking of the predictions.

5.2. Optimization via CMA-ES

The determination of the optimal weights w_i and temperature parameters T_i was formulated as a hyper-parameter optimization problem. We employed the CMA-ES (Covariance Matrix Adaptation

Evolution Strategy) algorithm [22] via the Optuna framework [23]. Unlike grid search, CMA-ES is a stochastic algorithm that models the parameter distribution through a covariance matrix, proving highly effective in continuous and non-linear search spaces. The objective function to be maximized over 3,500 trials is the Macro F1-Score computed on the validation set:

$$\mathcal{O} = \text{Macro-F1}\left(y_{val}, \text{argmax} \sum_{i=1}^{9} \bar{w}_i \cdot \frac{L_{i,val}}{T_i}\right) \tag{4}$$

subject to the constraint that the sum of the normalized weights must be equal to unity: $\sum_{i=1}^{9} \bar{w}_i = 1$, where $\bar{w}_i = \frac{w_i}{\sum_{j=1}^{9} w_j}$.

5.3. Final Inference Process

During the test phase, the framework aggregates the predictions according to the weighted scheme. As previously described in Section 4.5, for the Swin V2 5-patch model, the L_{patch} logits are pre-averaged at the image level before the global weighting. The final predicted class $\hat{y}$ is obtained as:

$$\hat{y} = \arg\max \left(\sum_{i=1}^{9} \bar{w}_i \cdot L'_i\right) \tag{5}$$

where $L'_i = \frac{L_i}{T_i}$ denotes the temperature-scaled logits of model i. This late fusion approach allows for the exploitation of both local and global specializations, ensuring superior robustness compared to individual models. The entire architecture is illustrated in Figure 4.

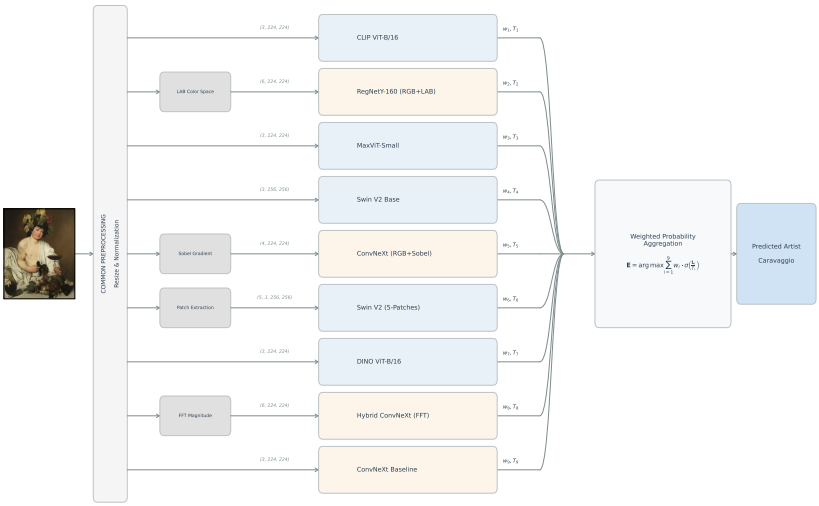

Figure 4: Overview of the Synthetic Multi-view Learning Late Fusion framework. The diagram illustrates the multi-perspective analysis process: the input image undergoes a common pre-processing stage and specific extraction branches. The nine heterogeneous models generate calibrated logits, which are finally combined via Weighted Probability Aggregation for the final artist prediction.

5.4. Result Analysis and Model Contribution

Optimization via CMA-ES enabled the identification of the optimal combination of weights and temperatures, as reported in Table 1. Analysis of the weights assigned by the optimization algorithm reveals several key insights. Although with varying magnitudes, all models contribute to the final Multi-View Learning prediction. The ConvNeXt 4-channel Structural model attained the highest weight (17.30%).

This result suggests that structural contours provide complementary discriminative information within the ensemble. The weight assigned to SwinV2 5-Patch (16.53%) highlights the importance of local textures, which are better preserved by operating on patches rather than on the full image. Additionally, the optimization procedure did not only evaluate the complete nine-model ensemble, but also tested sub-combinations ranging from two models upwards.

Table 1
Optimal weights and temperature parameters obtained through CMA-ES optimization.

Model	Final Weight (%)	Temperature (T)
CLIP (ViT-B/16)	6.83%	2.2274
DINO (ViT-B/16)	13.33%	1.4793
ConvNeXt Base	16.65%	2.2641
ConvNeXt Hybrid FFT	3.44%	1.4077
MaxViT Small	2.51%	2.4825
SwinV2 Base	11.47%	0.8840
SwinV2 5-Patch	16.53%	1.0006
RegNet 6ch (LAB)	11.93%	2.3305
ConvNeXt 4ch (Struct)	17.30%	1.4416

6. Experimental Results

In this section, we present the results obtained by the individual models and the Late Fusion framework on the test set. The evaluation was conducted using three primary metrics: Macro F1-Score, Balanced Accuracy, and Overall Accuracy.

6.1. Model Performance

Table 2 reports the test set results across Macro F1-Score, Balanced Accuracy, and Overall Accuracy. The Late Fusion framework significantly outperforms the best individual model (Swin V2 Base), boosting accuracy from 82.37% to 87.39%. In comparison with prior WikiArt literature, Zhao et al. [24] reached 93.50% using BiT ResNet-50 and Cetinic et al. [25] achieved 81.94% with fine-tuned CaffeNet, but both operated on a restricted subset of only 23 artists. In contrast, our approach scales effectively to 177 artists by fusing complementary multi-scale, chromatic, and structural representations.

Table 2
Performance comparison on the test set.

Model	Macro F1	Bal. Acc.	Over. Acc.
CLIP ViT-B/16	0.6916	0.6895	0.7063
DINO ViT-B/16	0.6875	0.6849	0.6995
MaxViT Small	0.7133	0.7113	0.7291
ConvNeXt Base	0.7475	0.7450	0.7611
Hybrid ConvNeXt (FFT)	0.7466	0.7430	0.7590
ConvNeXt 4-ch (Struct)	0.7622	0.7576	0.7711
RegNetY-160 (LAB)	0.7723	0.7702	0.7823
Swin V2 5-Patch	0.7795	0.7791	0.7917
Swin V2 Base	0.8124	0.8104	0.8237
Multi-view Learning	**0.8673**	**0.8611**	**0.8739**

6.2. Qualitative Analysis

To further evaluate the effectiveness of the proposed Late Fusion framework, Figure 5 showcases the model's predictions on a set of random samples from the test set, where the reported confidence is computed as the maximum softmax probability over the aggregated logits.

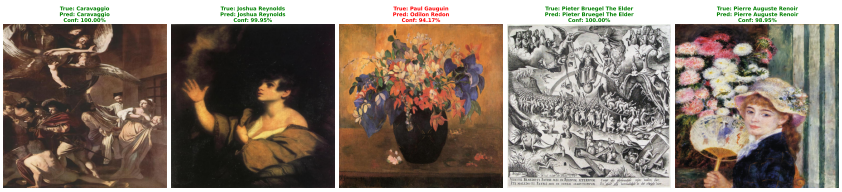

Figure 5: Visual inference examples. For each image, the true author (True), the predicted author (Pred), and the ensemble confidence level (Conf) are reported. Correct predictions are highlighted in green.

7. Conclusion

The introduced Multi-View Learning framework has successfully met the objectives established for artwork authorship attribution, demonstrating high accuracy and robustness in identifying an author's hand. However, despite these results, some limitations persist. The reliance on the WikiArt dataset introduces potential biases due to digitization inconsistencies, which might partially influence the learning of stylistic features rather than pure artistic technique. Furthermore, the multi-model architecture requires greater computational resources than single-model approaches, though this remains compatible with the specialized and non-instantaneous nature of art authentication. In terms of generalization, the system is currently trained on a closed set of 177 artists. While effective, it is fundamental to perform a fine-tuning process when analyzing specific work, focusing on the narrow circle of candidate artists relevant to a given analysis scenario. Looking ahead, future research aims to evolve this framework into a multimodal system integrating contextual domain knowledge, such as metadata related to support, period, and artistic technique. Furthermore, a potential evolution involves the integration of textual information from archival databases using Natural Language Processing techniques or Vision-Language architectures. Such an approach would allow for the exploitation of documented historical knowledge, steering the system toward a deeper semantic and historical understanding of the artworks.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Gemini for grammar, style, and clarity checks and sentence-level paraphrasing. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] V. Rathi, A. Sharma, A. Venkata Nithin, A. K. Singh, B. B. Gupta, A survey of computational techniques for fine art painting classification, Image and Vision Computing 161 (2025) 105626. URL: https://www.sciencedirect.com/science/article/pii/S0262885625002148. doi:https://doi.org/10.1016/j.imavis.2025.105626.

[2] B. Saleh, A. Elgammal, Large-scale classification of fine-art paintings: Learning the right metric on the right feature (2015). doi:10.11588/dah.2016.2.23376.

[3] L. Gatys, A. Ecker, M. Bethge, A neural algorithm of artistic style, Journal of Vision 16 (2016) 326. doi:10.1167/16.12.326.

[4] G. Ghiasi, H. Lee, M. Kudlur, V. Dumoulin, J. Shlens, Exploring the structure of a real-time, arbitrary neural artistic stylization network, 2017. doi:10.5244/C.31.114.

[5] F. Ji, M. Mcmaster, S. Schwab, G. Singh, L. Smith, S. e. a. Adhikari, Discerning the painter's hand: machine learning on surface topography, Heritage Science 9 (2021). doi:10.1186/s40494-021-00618-w.

[6] H. Yang, K. Min, A saliency-based patch sampling approach for deep artistic media recognition, Electronics 10 (2021) 1053. doi:10.3390/electronics10091053.

[7] C. Rodriguez, M. Lech, E. Pirogova, Classification of style in fine-art paintings using transfer learning and weighted image patches, 2018, pp. 1–7. doi:10.1109/ICSPCS.2018.8631731.

[8] G. Strezoski, M. Worring, Omniart: Multi-task deep learning for artistic data analysis (2017). doi:10.48550/arXiv.1708.00684.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. U. et al., An image is worth 16x16 words: Transformers for image recognition at scale, ArXiv abs/2010.11929 (2020). URL: https://api.semanticscholar.org/CorpusID:225039882.

[10] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. e. a. Wei, Swin transformer v2: Scaling up capacity and resolution, 2022, pp. 11999–12009. doi:10.1109/CVPR52688.2022.01170.

[11] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11966–11976. doi:10.1109/CVPR52688.2022.01167.

[12] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, 2021, pp. 9630–9640. doi:10.1109/ICCV48922.2021.00951.

[13] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. e. a. Agarwal, Learning transferable visual models from natural language supervision, arXiv preprint arXiv:2103.00020 (2021). doi:10.48550/arXiv.2103.00020.

[14] W. Fedus, B. Zoph, N. Shazeer, Switch transformers: scaling to trillion parameter models with simple and efficient sparsity, J. Mach. Learn. Res. 23 (2022). doi:10.5555/3586589.3586709.

[15] Z. Yu, Z. Dong, C. Yu, K. Yang, Z. Fan, C. Chen, A review on multi-view learning, Frontiers of Computer Science 19 (2024). doi:10.1007/s11704-024-40004-w.

[16] M. Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. e. a. Singhal, Fourier features let networks learn high frequency functions in low dimensional domains, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020. doi:10.5555/3495724.3496356.

[17] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. e. a. Hamprecht, On the spectral bias of neural networks, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 5301–5310. URL: https://proceedings.mlr.press/v97/rahaman19a.html.

[18] A. Avanaki, K. Espig, T. Kimpe, A. Xthona, J. Rostang, B. Piepers, Perceptual uniformity of commonly used color spaces, Proceedings of SPIE - The International Society for Optical Engineering 9041 (2014). doi:10.1117/12.2043524.

[19] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, P. Dollar, Designing network design spaces, 2020, pp. 10425–10433. doi:10.1109/CVPR42600.2020.01044.

[20] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, Y. Li, MaxViT: Multi-axis Vision Transformer, 2022, pp. 459–479. doi:10.1007/978-3-031-20053-3_27.

[21] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, JMLR.org, 2017, p. 1321–1330. doi:10.5555/3305381.3305518.

[22] N. Hansen, A. Ostermeier, Completely derandomized self-adaptation in evolution strategies, Evol. Comput. 9 (2001) 159–195. URL: https://doi.org/10.1162/106365601750190398. doi:10.1162/106365601750190398.

[23] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 2623–2631. URL: https://doi.org/10.1145/3292500.3330701. doi:10.1145/3292500.3330701.

[24] W. Zhao, W. Jiang, X. Qiu, Big transfer learning for fine art classification, Computational Intelligence and Neuroscience 2022 (2022) 1–19. doi:10.1155/2022/1764606.

[25] E. Cetinic, T. Lipic, S. Grgic, Fine-tuning convolutional neural networks for fine art classification, Expert Systems with Applications 114 (2018). doi:10.1016/j.eswa.2018.07.026.