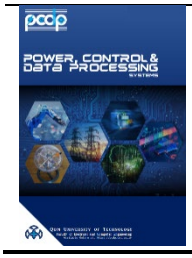
DOI: 10.30511/pcdp.2026.2095440.1101 - ISSN: 3060-8376 - Vol 4(1), 2027



POWER, CONTROL AND DATA PROCESSING SYSTEMS

Available Online at: https://pcdp.qut.ac.ir/

Classification of Persian Miniature Painting Schools Using a Vision Transformer and Transfer Learning

ARTICLE INFO

Article Type
Original Research

Authors
Darya Seifi Zarei[1]
Hamid Reza Shahdoosti[2]

[1] Department of electrical and computer engineering, Hamedan University of Technology, Hamedan, Iran, darya.seifi@stu.hut.ac.ir
[2] Department of Electrical and computer engineering, Hamedan University of Technology, Hamedan, Iran, h.doosti@hut.ac.ir

**** Correspondence***
Address: Department of electrical and computer engineering, Hamedan University of Technology, Pajohesh Square, Hamedan, Iran. Postal Code: 65155579.
Phone: 988138411529
Fax: 988138380660

h.doosti@hut.ac.ir

Article History
Received: March 16, 2022
Accepted: June 09, 2022
ePublished: September 20, 2022

ABSTRACT

The automated attribution of historical paintings to their school of origin is a long-standing challenge at the intersection of computer vision and art history. This challenge is especially acute for non-Western art traditions, such as Persian miniature painting, for which digitized and labelled image corpora remain comparatively small. This paper addresses the classification of Persian miniature paintings into five historical schools including, Herat, Qajar, Shiraz-e Avval, Tabriz-e Avval, and Tabriz-e Dovvom, using a Vision Transformer (ViT). Transfer learning is used to adapt a ViT-B/16 model, pre-trained on ImageNet-21k, to this task. Because the available dataset contains only 890 images, five complementary augmentation operations i.e. random-resized cropping, horizontal flipping, rotation, color jitter, and random erasing, are applied during training, together with Mixup as an additional sample-mixing regularizer. A progressive layer-freezing strategy keeps the first six Transformer encoder blocks fixed and fine-tunes only the last six blocks. To stabilize training on the small fine-tuning set, this strategy utilizes a classification head combined with label smoothing and a low, cosine-annealing learning rate. On a held-out test set of 178 images, the fine-tuned model reaches an overall accuracy of 92.13% and a weighted F1-score of 0.92, with three of the five schools classified at an F1-score above 0.94. These results indicate that a Vision Transformer, can reach competitive accuracy on a small, specialized cultural-heritage dataset when transfer learning and augmentation are used simultaneously.

Keywords: Vision Transformer; Persian miniature painting; transfer learning; data augmentation; image classification; digital humanities.

1 Introduction

In recent years, deep-learning-based image analysis has advanced rapidly and has been applied to a wide range of historical and artistic materials, including manuscript and calligraphy recognition, condition assessment of degraded artworks, and the study of archaeological structures and artifacts [1-3]. Motivated by this progress, the present study applies a deep-learning framework to Persian miniature painting, a rich but computationally underexplored part of Iranian visual heritage.

Persian miniature painting developed over more than five centuries under successive royal ateliers. The Herat school flourished under the late Timurids; the Tabriz and Shiraz schools developed across the early and middle Safavid period; and the Qajar court workshops produced a further, visually distinct tradition in the nineteenth century [4, 5]. Each school is recognizable not only through subject matter but through a coherent set of pictorial conventions, including, color palette, compositional density, the treatment of rock and cloud forms, headgear silhouette, and the quality of contour lines. Correct school attribution therefore carries real value for provenance research, conservation planning, and the cataloguing of manuscript folios dispersed across museums and private collections.

Classifying paintings by school is difficult for a computational model because the differences between schools are often subtle and localized (headdress shape, particular rock formation and narrow band of color) in subject matter [5]. Hand-crafted image descriptors, such as color histograms and texture statistics, have historically struggled to capture this kind of fine-grained stylistic detail, which is one of the reasons that hierarchical representations have become the default approach for painting classification in recent years. Convolutional Neural Networks (CNNs) have so far been the dominant architecture for this task, owing to a built-in assumption that nearby pixels are related and that useful visual patterns recur across spatial locations [6].

A further practical constraint in this domain is the limited size of most cultural-heritage image collections. Large-scale digitization efforts exist for a small number of well-resourced Western archives, but comparable resources are rare for many other pictorial traditions, Persian miniature painting included. Transfer learning — initializing a network with weights learned on a large, general purpose image corpus and adapting it to a small, specialized dataset — is one of the principal strategies used to address this constraint, and it is the strategy adopted in this study [7, 8]. Unlike traditional convolutional filters that are restricted to local neighborhoods, Vision Transformers utilize a self-attention mechanism to relate all image regions from the very first layer. This capability has enabled ViTs to achieve state-of-the-art performance across various image classification and recognition tasks [8]. This demonstrated capability motivated the use of a ViT architecture in the present study, in place of the convolutional networks that have so far been the default choice for painting classification.

The remainder of this paper is organized as follows. Section 2 reviews related work on computational painting classification. Section 3 describes the dataset. Section 4 presents the proposed methodology, including the model architecture, the training protocol, and the evaluation metrics. Section 5 reports and discusses the experimental results. Section 6 concludes the paper.

2 Related Work

Automated painting classification was initially performed using classical machine learning approaches. Early work combined hand-crafted color and edge descriptors with decision-tree classifiers to sort digitized paintings into broad genres, showing that the task was feasible even without learned features [9]. Later studies employed Support Vector Machines (SVM) and evolutionary-optimized k-nearest-neighbor classifiers for artist attribution on small, single-collection datasets, improving accuracy by fusing texture, color, and geometric descriptors rather than relying on any single feature type [10, 11]. The introduction of learned, hierarchical representations substantially changed this landscape: a convolutional autoencoder used purely for feature extraction, paired with a shallow classifier, was shown to outperform hand-crafted pipelines by a wide margin on painter attribution [12]. Building on this direction, patch-based approaches that divide each painting into regions, classify each region independently with a pre-trained CNN, and fuse the resulting predictions were proposed for large, multi-style art collections, and this fusion strategy substantially improved over single-pass whole-image classification [13].

Within Persian visual culture specifically, computational classification has received comparatively little attention. One of the few examples combined hand-crafted image-processing features — obtained through segmentation and edge- and shape-based descriptors — with an SVM to automatically grade Persian Nastaliq calligraphy strokes against a reference style [14]. This feature-engineering burden mirrors the broader limitation that motivated the shift toward learned, hierarchical representations elsewhere in the literature.

Vision Transformers have more recently been applied to Western painting style classification on large, well-digitized corpora such as WikiArt, reaching competitive accuracy and suggesting that global self-attention captures compositional relationships [15]. Hierarchical attention variants that restrict self-attention to local windows for computational efficiency have also been evaluated on the same task, but with a noticeable drop in accuracy, indicating that the genuinely global form of attention is what benefits the problem of art style classification [16]. Hybrid architectures that combine convolutional and Transformer components within a single network have been proposed on the premise in which local and global features are complementary [17]. Subsequently, Wang

and Huang [18] investigated several neural architectures, such as CNNs, GANs, and ViTs, to distinguish between various art styles, including abstract, realist, impressionist, and digital forms. Their findings demonstrated that hybrid CNN+ViT configurations provide a robust foundation for both museum curation and computational creativity. Zhang and Lv [19] proposed a hybrid model that integrates a Vision Transformer with a lightweight CNN through a multi-scale attention-weighted fusion strategy, significantly outperforming standalone baselines by effectively combining fine-grained local details with global contextual understanding. Similarly, Wang and Song [20] developed a fusion model that first extracts local features with a CNN, then captures global dependencies with a Transformer, and finally fuses both representations. Experiments on both Chinese and oil painting datasets showed that this combined approach consistently improved classification accuracy over individual CNN or Transformer models alone, demonstrating the value of complementary feature extraction for artistic media.

Beyond style recognition, ViT-based models have been actively applied to historical documents and broader cultural heritage material. Parres and Paredes [21] fine-tuned a vision encoder-decoder architecture built around a ViT backbone for transcribing handwritten historical manuscripts. Aided by a dedicated augmentation and fine-tuning strategy, their system outperformed earlier convolutional-recurrent pipelines on several benchmark collections, highlighting the importance of adapting transformers to document analysis. Roald et al. [22] presented a proof-of-concept image retrieval and classification system for digitized pre-1900 books from the National Library of Norway and compared ViT, CLIP (Contrastive Language-Image Pre-Training), and SigLIP (Sigmoid loss for Language-Image Pre-training) architectures. Sivan and Pati [23] automated the classification of palm-leaf heritage documents into visual quality categories by integrating Discrete Cosine Transform features with a Vision Transformer, achieving robust performance. This research demonstrated that incorporating classical transform-domain information can strengthen a Transformer-based model when training data is scarce. Vidal-Gorène and Decours-Perez [24] tackled the detection and interpretation of damaged medieval Armenian inscriptions using a two-stage pipeline. In their work, YOLO (You Only Look Once) was used for localization and a Vision Transformer was used for character classification.

A separate line of work addresses the practical difficulty of training Transformers when labelled data is scarce more generally. Research on artist authentication has shown that Swin Transformers can achieve high discriminative performance even on challenging imitation-only sets [25]. This method outperforms convolutional models, suggesting that ViTs are particularly well-suited for detecting subtle forgeries where global compositional coherence matters more than local texture. In the domain of archaeological ceramics, Capriotti et al. [26] compared CNNs and ViTs directly for classifying petrographic thin-sections. They showed that transfer learning substantially improved both architectures, though the ViT benefited more from augmentation and obtained better performance. This observation aligns with broader studies on data-efficient training of Transformers, where strategies such as knowledge distillation from a convolutional teacher or aggressively applying data augmentation have all been used to reduce reliance on massive pre-training corpora [27, 28]. In parallel, data augmentation has itself become an active research area within cultural-heritage computing. Recent work has explored using diffusion models to generate synthetic art samples for style classification, showing that augmenting real training sets with high-quality synthetic images can materially improve generalization and reduce overfitting. Attention mechanisms and explainability tools have also been increasingly integrated into art analysis pipelines, with Grad-CAM (Gradient-weighted Class Activation Mapping) and attention-rollout methods being applied to artist and style classification models to visualize which regions of a painting drive prediction [29].

Taken together, these studies indicate that ViT's competitiveness on small, specialized image collections depends considerably on whether the training protocol actively compensates for the limited amount of data through augmentation, progressive freezing, or hybridization. However, none of the above approaches has previously been evaluated on non-Western art traditions such as Persian miniature painting, nor on a dataset as small as the 890-image corpus considered here. Moreover, no prior study has provided an interpretability analysis linking a Transformer's internal attention patterns to art-historically meaningful features within this specific domain. The present work addresses this gap by combining the transfer learning, a progressive layer-freezing strategy, and a comprehensive data augmentation regime such as Mixup to adapt a ViT-B/16 [30] model to the five-school Persian miniature classification task. Our work also offers attention-based analyses that connect model behavior to documented artistic lineages.

3 Dataset

The dataset used in this study was assembled and introduced together with a convolutional baseline in a prior publication [31] and is used here without modification, so that results remain directly comparable. It comprises 890 high-resolution digitized images (a minimum of 1000 pixels on the longer side), drawn from the Islamic Painted Page database and from specialized archives — including the National Library of Iran, the Reza Abbasi Museum, the Library of the Islamic Consultative Assembly, the Museum of Contemporary Art Tehran, the Metropolitan Museum of Art, and the Victoria and Albert Museum — and labelled by a specialist in Islamic art history under expert supervision. Images that were ambiguous in school attribution or of insufficient photographic quality were excluded, and extraneous margins were manually cropped so that only pictorially relevant content remains. Table 1 summarizes the class distribution and the stratified

80/20 train–test split, which preserves per-class proportions in both partitions given the moderate class imbalance (150–202 images per school).

Table 1: Class distribution and stratified train/test split of the Persian Miniature dataset.

Artistic School	Historical Period	Training Images	Test Images	Total
Herat	Timurid era (15th c.)	140	40	180
Qajar	Qajar dynasty (19th c.)	170	30	200
Shiraz-e Avval	Late Timurid – early Safavid	121	39	160
Tabriz-e Avval	Early Safavid (16th c.)	132	38	170
Tabriz-e Dovvom	Mid-Safavid (16th c.)	149	31	180
Total	—	**712**	**178**	**890**

Figure 1 shows one representative image from each of the five schools. The Qajar school (Figure 1, second panel) departs sharply from earlier convention through European-influenced portraiture and chiaroscuro shading.



a b c d e

Figure 1. Representative examples from the five Persian miniature painting schools studied in this work: (a) Herat, (b) Qajar, (c) Tabriz-1, (d) Shiraz-1, and (e) Tabriz-2.

Shiraz-e Avval favors a comparatively sparse composition dominated by red. Tabriz-e Avval and Tabriz-e Dovvom share a dense, richly ornamented compositional style with the Herat school, from which both descend, which is the principal source of visual overlap discussed later in Section 5.

The dataset is modest in size relative to the datasets typically used to train or fine-tune Vision Transformers. Typically, deep networks are conventionally pre-trained on tens of millions of images and fine-tuned on downstream sets of several thousand images or more. Only a few images are available here for fine-tuning, because all our training data is 712 images that should be used for training and validation. This scarcity is not specific to Persian miniatures. It is characteristic of most cultural-heritage image collections, which makes the dataset a realistic test of how well a Transformer-based model can be adapted to a genuinely small, specialized corpus.

4 Methodology

4.1 Preprocessing

Each image is resized so that its longer side does not exceed 256 pixels, preserving the original aspect ratio. Any resulting empty area is padded symmetrically to produce a square canvas, and a center crop to 224 × 224 pixels is then applied, matching the fixed input resolution required by the ViT patch-embedding layer. Pixel values are normalized using the ImageNet channel-wise mean and standard deviation rather than statistics computed on the Persian miniature corpus itself. Because the ViT-B/16 backbone was pre-trained on ImageNet, reusing its original normalization statistics avoids a mismatch between the low-level pixel statistics expected by the pre-trained weights and those present in the fine-tuning data, which was found to be important for stable convergence. Figure 2 summarizes this pipeline schematically, from the original image to the normalized tensor (RGB format of the 224× 224 image) passed to the model.

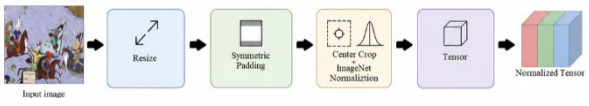


Figure 2: Preprocessing pipeline applied to every image before it is passed to the ViT-B/16 model.

4.2 Data augmentation

Given the limited size of the fine-tuning set, five augmentation operations are applied to every training image, independently and stochastically, at every epoch: (1) random-resized cropping (scale 0.80–1.00), simulating variation in zoom and framing; (2) random horizontal flipping (p = 0.5), admissible here because Persian miniature composition carries no fixed left–right convention; (3) random rotation within ±15°, improving robustness to imperfect page alignment in manuscript photography; (4) color jitter of brightness, contrast, saturation, and hue, encouraging reliance on

structural and relative-color patterns rather than absolute pigment intensity, which varies with scanning conditions and manuscript age; and (5) random erasing ($p = 0.2$), discouraging reliance on any single, spatially fixed cue. Because these five operations are re-sampled independently at every training epoch, the model is exposed to well over 42720 distinct augmented views of the 712 training images over the course of the 60-epoch training schedule.

In addition to these five operations, Mixup [32] is applied with $\alpha = 1.0$. Mixup trains the network on convex combinations of pairs of training images and their one-hot labels, rather than on individual images alone, and is applied to the training set only. Note that, the training data and the test set are never mixed. Because Mixup smooths the decision boundary between classes. It is particularly useful for schools such as Herat and Tabriz-e Dovvom, whose stylistic overlap (discussed in Section 5.3). Figure 3 shows the effect of the five geometric and photometric operations on a representative training image.



Figure 3: Effect of the five augmentation operations applied during training, shown on a representative Herat-school image.

4.3 Model Architecture

The classification model is ViT-B/16 (Vision Transformer, Base configuration, 16 × 16 patches), initialized from weights pre-trained on ImageNet-21k (≈ 14.2 million images). Table 2 summarizes the architectural configuration, and Figure 4 illustrates the overall data flow.

Table 2: Architectural configuration of the ViT-B/16 model used in this study.

Parameter	Value
Total parameters	≈ 86.0 million
Transformer encoder blocks	12
Attention heads per block	12
Hidden dimension	768
MLP dimension	3072 (4 × 768)
Patch size	16 × 16 pixels
Patches per 224 × 224 image	196 (14 × 14)
Input sequence length	197 (196 patch tokens + 1 classification token)
Pre-training corpus	ImageNet-21k (≈ 14.2 million images)
Dropout before classifier	0.30 (raised from the pre-training default of 0.10)

Each 224 × 224 input image is partitioned into 196 non-overlapping 16 × 16 patches, each linearly projected to a 768-dimensional embedding. A learnable classification token is prepended to the resulting sequence, and learnable position embeddings are added, since the subsequent self-attention operation is otherwise insensitive to patch ordering. The 197-token sequence is then processed by 12 Transformer encoder blocks, each combining multi-head self-attention (12 heads) with a position-wise feed-forward network (hidden dimension 3072), residual connections, and layer normalization around both sub-layers. The final classification token representation is passed to a linear classification head with five outputs, one per school, preceded by a dropout layer (rate 0.30) that guards against overfitting on the comparatively small fine-tuning set.

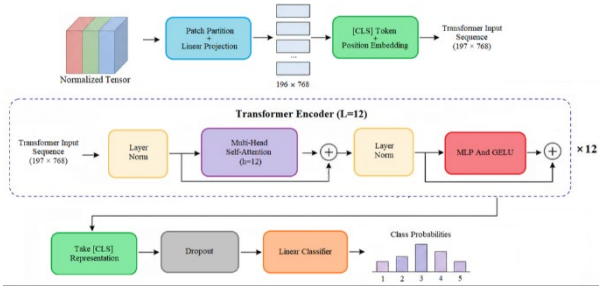


Figure 4: Overview of the ViT-B/16 architecture used in this study, from patch extraction to the final classification head. Blocks 1–6 of the Transformer encoder is kept frozen during fine-tuning, while blocks 7–12 and the classification head are updated.

4.4 Transfer Learning and Fine-Tuning Strategy

Fine-tuning all 86 million parameters directly on the training images risks erasing the general-purpose visual representations learned during ImageNet-21k pre-training. A progressive, partial layer-freezing strategy is therefore adopted: the first six of the twelve encoder blocks are kept frozen throughout training, preserving low-level, broadly transferable features (edges, textures, simple color contrasts), while the deeper six blocks and the classification head are fine-tuned. Freezing roughly half of the encoder also reduces the number of trainable parameters relative to the size of the fine-tuning set, acting as a computationally free form of regularization. The classification head is randomly re-initialized, since the task changes from 1000-way ImageNet classification to five-way school attribution.

Table 3 summarizes the full training configuration. Here, the AdamW optimizer [33] is used for training. This optimizer decouples weight decay from the gradient update, which has been shown to transfer more reliably to Transformer architectures than the standard Adam. To prevent destructive updates to the pre-trained weights, a low initial learning rate of 1×10^{-5} was utilized. A cosine-annealing schedule decays the learning rate smoothly toward zero over training, and early stopping with a patience of 10 epochs, monitored on a held-out validation split (15% of the training data), selects the final

checkpoint. To prevent data leakage, data augmentation was performed at the training-set level.

Table 3: Training hyperparameters for ViT-B/16 fine-tuning.

Hyperparameter	Value
Initial learning rate	1×10^{-5}
Optimizer	AdamW [33]
Weight decay	0.03
Learning-rate schedule	Cosine annealing, Tmax = 60
Loss function	Cross-entropy with label smoothing (ε = 0.10)
Batch size	32
Maximum epochs	60
Early-stopping patience	10 epochs, monitored on a held-out validation split
Layer-freezing schedule	Encoder blocks 1–6 frozen; blocks 7–12 and the classification head fine-tuned
Augmentation (training only)	Random-resized crop, horizontal flip, ±15° rotation, color jitter, random erasing, plus Mixup (α = 1.0)

4.5 Evaluation metrics

Model performance is reported using overall accuracy together with per-class precision, recall, and F1-score, and their macro- and weighted averages. Table 4 lists the corresponding formulas. In Table 4, TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. In addition, C is the total number of classes (C = 5 in this study). The overall accuracy is reported in Section 5.1 for the purpose of comparing different methods. Per-class precision, recall, and F1 are also reported because accuracy alone can obscure class-specific failure modes in a moderately imbalanced five-class setting.

Table 4: Evaluation metrics used in this study.

No.	Formula
(1)	$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$
(2)	$\text{Precision}_c = \frac{TP_c}{TP_c+FP_c}$
(3)	$\text{Recall}_c = \frac{TP_c}{TP_c+FN_c}$
(4)	$F1_c = \frac{2\times Precision_c \times Recall_c}{Precision_c + Recall_c}$
(5)	$\text{Macro0F1} = \frac{1}{C}\sum_{c=1}^{C} F1_c$

5 Results and discussion

5.1 Classification Accuracy

A 10-fold validation approach was designed through repeated random train-validation splits. As explained in section 4.4, the training partition was subdivided into an 85% training set and a 15% validation set to optimize hyperparameters. The 712-image development set was repeatedly divided into 605 training images and 107 validation images, while preserving the class distribution across the two subsets. To prevent data leakage, no augmented version of an image represented in the training set was present in the validation set. The model consistently achieved a mean accuracy of 97.4% with a standard deviation (SD) of 1.5% for the validation set, demonstrating high stability across the folds. Table 5 places this result alongside the results of powerful CNN-based networks including, DenseNet201 [34] and DenseNet169 [35]. In addition, the fine-tuned ViT-B/16 model reaches an overall accuracy of 92.13% and a weighted F1-score of 0.920 on the held-out test set of 178 images. While the DenseNet201 demonstrated performance levels close to our proposed approach, it exhibited greater variability across the evaluated splits.

Table 5: Comparison of different methods in classifying Persian painting schools.

Model	Accuracy for validation data (%)	Accuracy for test data (%)
DenseNet201 [34]	96.77± 3.2%	91.57
DenseNet169 [35]	94.36± 2.9%	89.89
This study	**97.40 ± 1.5%**	**92.13**

5.2 Per-class performance

Table 6 and Figure 5 report the full per-class classification report and confusion matrix on the 178-image test set. Three schools (Qajar, Shiraz-e Avval, and Tabriz-e Avval) are classified with an F1-score at or above 0.94, and Shiraz-e Avval is recognized with perfect recall.

Table 6: Per-class classification report for ViT-B/16 on the held-out test set, derived from the confusion matrix in Figure 5.

Artistic School	Precision	Recall	F1-score	Support
Herat	0.857	0.900	0.878	40
Qajar	0.935	0.967	0.951	30
Shiraz-e Avval	0.951	1.000	0.975	39
Tabriz-e Avval	0.947	0.947	0.947	38
Tabriz-e Dovvom	0.923	0.774	0.842	31
Macro average	0.923	0.918	0.919	178
Weighted average	0.921	0.921	0.920	178

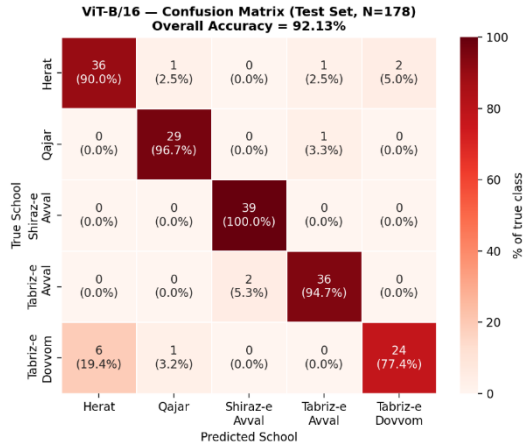


Figure 5: Confusion matrix of ViT-B/16 predictions on the 178-image held-out test set. Cell values show raw counts with the corresponding row-wise (recall) percentage.

This result is consistent with the strong visual distinctiveness of these schools: Qajar painting departs sharply from earlier

convention through chiaroscuro and frontal, European-influenced portraiture, while Shiraz-e Avval is characterized by a comparatively sparse composition and a dominant red with few analogues among the other four schools.

5.3 Confusion Among Historically and Stylistically Related Schools

The weakest class is Tabriz-e Dovvom (F1 = 0.842, recall = 0.774), and its errors are concentrated almost entirely in one direction: 6 of the 31 true Tabriz-e Dovvom test images (19.4%) are misclassified as Herat, while confusions in the opposite direction are rare. This finding may be interpreted in light of the historical relationship between the two traditions. In particular, the migration of Kamāl al-Dīn Behzād, a major figure associated with the Herat school, to Tabriz and his subsequent influence on artists working in the Tabriz tradition may have contributed to the transmission and transformation of stylistic features between the two schools [36]. Artists trained in this context may have retained elements of the Herat tradition while incorporating features acquired from the Tabriz artistic environment, potentially resulting in works that share visual characteristics with both traditions. However, this historical relationship should be regarded as a plausible art-historical interpretation rather than a demonstrated causal explanation for the classification errors observed in the present study.

5.4 Attention based interpretability

A recurring concern with deep classifiers in cultural-heritage applications is that high accuracy alone does not establish that a model's decisions are grounded in art-historically meaningful evidence, rather than incidental correlations such as scanner artifacts or archive-specific photographic conventions. This concern is particularly relevant for ViT, whose self-attention mechanism can in principle relate any two patches regardless of spatial distance. Two complementary analyses are reported below: a localization analysis using Grad-CAM, and a layer-wise analysis of how broadly the classification token distributes its attention across the encoder. Grad-CAM [37] was computed with respect to the predicted-class logit, using the LayerNorm preceding self-attention in the final encoder block as the target layer. The resulting 196-token relevance map was reshaped to its native 14 × 14 spatial grids, upsampled, and overlaid on the source image. Figure 6 shows the result for four Herat-school test images spanning different compositions including, a crowded battle scene, an architectural courtyard scene, a mounted night scene, and an outdoor genre scene. The model consistently prioritizes diagnostic features over non-informative regions. High-relevance areas are concentrated on figures and their attire rather than on peripheral elements such as borders or empty parchment. This is evident in the courtyard and night scenes, where the attention shifts away from architectural framing and mounted horses toward the core subjects. By leveraging Grad-CAM analysis, we can effectively visualize the regions of interest within the input images, thereby clarifying the features that drive the network's classification performance.

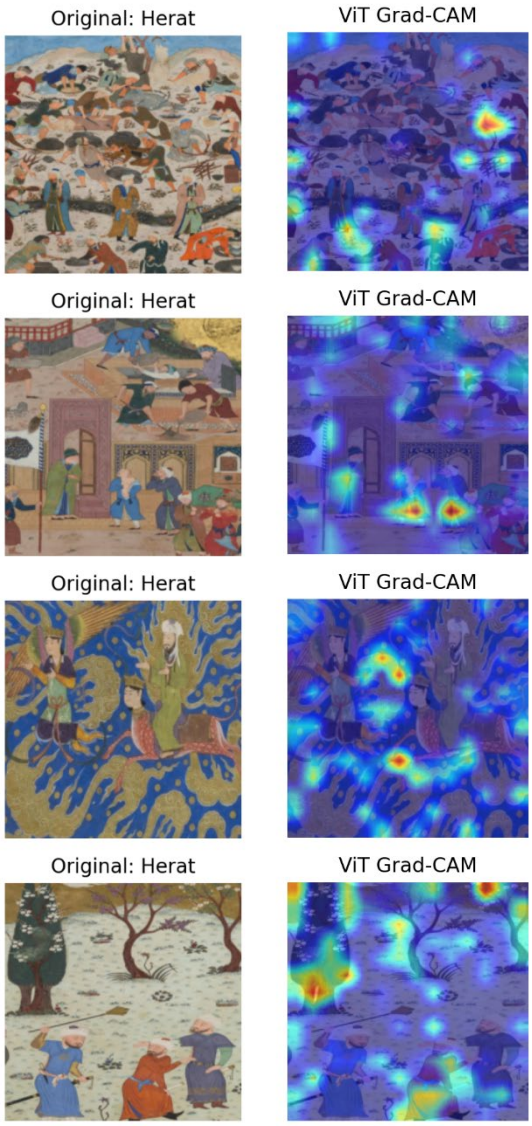


Figure 6: Grad-CAM localization for ViT-B/16 on four Herat-school test images, computed with respect to the predicted class and the final encoder block.

The second analysis quantifies how broadly the classification token's attention is spread across the 196 patch tokens. To this end, we use the Shannon entropy of attention distribution at each encoder block (averaged over a sample of 20 validation images). A low entropy indicates that the classification token concentrates on a small number of patches at that layer. In contrast, a high entropy indicates that it draws on evidence spread across most of the image. Figure 7 shows that entropy remains low and relatively unstable across the six frozen encoder blocks. Specifically, entropy starts at approximately 1.8 nats at block 1, rises unevenly to about 2.9 nats by block 4, and subsequently decreases. This pattern is consistent with these blocks retaining the more locally-focused attention

patterns learned during ImageNet pre-training. From block 6 onward — the point at which fine-tuning begins — entropy rises consistently, reaching its highest values in the last two blocks (approximately 4.3–4.4 nats at blocks 11 and 12). This indicates that fine-tuning shifted the classification token toward integrating evidence from across the entire patch grid by the final layers, rather than continuing to rely on a small number of locally salient patches. This behavior is consistent with the nature of the classification task, where school attribution depends on the joint configuration of spatially distant feature (such as headgear, background architecture, and palette). The increasing breadth of attention in the fine-tuned blocks provides a mechanistic account of how the network integrates this distributed evidence into a single prediction, thereby complementing the localized, patch-level evidence shown in Figure 6.

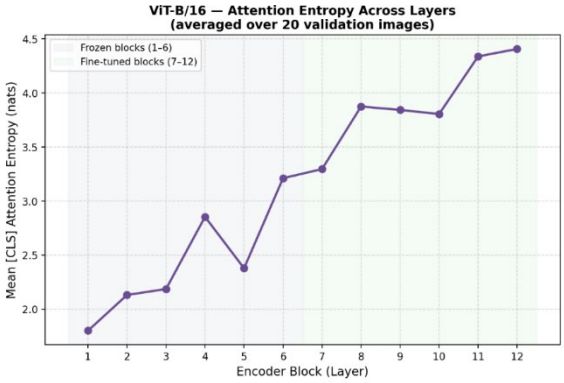


Figure 7: Mean entropy of the classification token's attention distribution at each of the 12 ViT-B/16 encoder blocks, averaged over 20 validation images. Shaded regions distinguish the frozen (blocks 1–6) and fine-tuned (blocks 7–12) portions of the encoder.

5.5 Limitation and Future Work

The dataset, while carefully curated, remains modest in size (890 images), and the test partition for the weakest class (31 images) is small enough that individual misclassifications noticeably affect the reported per-class metrics. Several directions can be pursued to extend the present work. First, the dataset can be enlarged with additional images and schools, which would strengthen the learned representations and reduce the reliance on data augmentation. Second, the frequent misclassification of Tabriz-e Dovvom as Herat suggests that the stylistic relationship between closely related schools deserves further investigation, possibly through finer-grained or hierarchical classification. Third, alternative Transformer architectures and art-specific pre-training strategies may be compared with ViT-B/16 to assess their benefit over ImageNet-based transfer learning. Fourth, the interpretability analyses based on Grad-CAM and attention entropy can be extended into a formal evaluation of model faithfulness. Finally, the generalization of the proposed model could be examined through cross-dataset evaluation on other collections of Iranian art. It is important to note that a detailed ablation study was not performed as the proposed method relies on the synergistic combination of these specific techniques (five augmentation operations, Mixup, label smoothing, dropout, partial freezing, and cosine annealing). These components are functionally interdependent; therefore, the removal of any individual element results in a significant degradation in model performance, rendering an isolated evaluation of their contributions inapplicable.

6 Conclusion

This paper presented, to the authors' knowledge, the first application of a Vision Transformer to the classification of Persian miniature painting schools as well as Iranian artworks. The proposed protocol combined transfer learning from a large, general-purpose pre-training corpus with five complementary data augmentation techniques and a progressive layer-freezing fine-tuning strategy, jointly addressing the limited size of the available dataset. The resulting ViT-B/16 model reached an overall accuracy of 92.13% and a weighted F1-score of 0.920 on a held-out test set, a result that is directly competitive with powerful deep neural networks architectures such as DenseNet201 and DenseNet169. This outcome indicates that the combined use of transfer learning and data augmentation can successfully mitigate the small size of the training set, allowing a Transformer-based architecture to classify Persian miniature painting schools with high accuracy. To enhance the interpretability of the proposed model, we employed Grad-CAM analysis. This allows us to gain deeper insights into the network’s decision-making behavior and to visualize the specific image features that are most salient to the model’s predictions.

Disclosure of Potential Conflicts of Interest

The Authors declare that there is no conflict of interest.

Reference

[1] Huang, Q., Li, M., Agustin, D., Li, L. and Jha, M., (2023). A novel CNN model for classification of Chinese historical calligraphy styles in regular script font. Sensors, 24(1), 188-197. https://doi.org/10.3390/s24010197

[2] Ghasemi, S., Ghasemi, S., Rahmani, A. M., & Barzamini, R. (2025). Transfer Learning-based Detection of COVID-19 Cases from Chest CT Scans. *Power, Control, and Data Processing Systems*, *2*(3), 33-47. https://doi.org/10.30511/pcdp.2025.2063435.1030

[3] Amirbeigi, H., Shahdoosti, H. R., (2026) A New Method for Detecting Emotions in Video Using a Hybrid CNN-RNN-RBM Network. *Power, Control, and Data Processing Systems*, (under publication). https://doi.org/10.30511/pcdp.2026.2087714.1069

[4] Jamali, L., 2013. Colouring the words: Reproduction of persian literature through the art of Iranian miniature. Literature and Interarts: Critical Essays, 29-52. https:// https://publicaciones.unirioja.es/catalogo/online/interarts/pdf/02_Jamali.pdf

[5] Rajabi, Z. and Hassanvand, M.K., (2022). An Investigation into Metaphorical Function in Human Forms of Persian Painting of Iran (Herat and Second Tabriz School). Painting Graphic Research, 4(7), 41-54. https://doi.org/10.22051/pgr.2022.40634.1138

[6] Sandoval, C., Pirogova, E. and Lech, M., (2019). Two-stage deep learning approach to the classification of fine-art paintings. IEEE access, 7, 41770-41781. https://doi.org/10.1109/access.2019.2907986

[7] Iman, M., Arabnia, H.R. and Rasheed, K., (2023). A review of deep transfer learning and recent advancements. Technologies, 11(2), 29-40. https://doi.org/10.3390/technologies11020040

[8] Yao, S., Kang, Q., Zhou, M., Rawa, M.J. and Abusorrah, A., (2023). A survey of transfer learning for machinery diagnostics and prognostics. Artificial Intelligence Review, 56(4), 2859-2871. https://doi.org/10.1007/s10462-022-10230-4

[9] Zujovic, J., Gandy, L., Friedman, S., Pardo, B., Pappas, T.N., (2009). Classifying paintings by artistic genre: An analysis of features & classifiers. In: 2009 IEEE International Workshop on Multimedia Signal Processing, 1-5. https://doi.org/10.1109/mmsp.2009.5293271

[10] Song, H., Wang, L., & Song, C. (2025). Research on the analysis of image characteristics in fine art painting under the application of machine learning technology. Ain Shams Engineering Journal, 16(8), 103456. https://doi.org/10.1016/j.asej.2025.103456

[11] Levy, E., David, O.E., Netanyahu, N.S., (2014). Genetic algorithms and deep learning for automatic painter classification. In: Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation, 1143-1150. https://doi.org/10.1145/2576768.2598287

[12] David, O.E., Netanyahu, N.S., (2016). DeepPainter: Painter classification using deep convolutional autoencoders. In: Artificial Neural Networks and Machine Learning – ICANN 2016, Part II. Springer, 20-28. https://doi.org/10.1007/978-3-319-44781-0_3

[13] Sandoval, C., Pirogova, E., Lech, M., (2019). Two-stage deep learning approach to the classification of fine-art paintings. IEEE Access, 7, 41770-41781. https://doi.org/10.1109/access.2019.2907986

[14] Yadav, R. K., Yadav, S. K., Yadav, M., Yadav, R., Shankar, A., & Amoon, M. (2025). Classification and feature extraction of text from hindi document for optical character recognition: RK Yadav et al. International Journal on Document Analysis and Recognition (IJDAR), 1-27. https://doi.org/10.1007/s10032-025-00552-8

[15] Iliadis, L.A., Nikolaidis, S., Sarigiannidis, P., Wan, S. and Goudos, S.K., (2021). Artwork style recognition using vision transformers and mlp mixer. Technologies, 10(1), 12-24. https://doi.org/10.3390/technologies10010002

[16] L Zhang, Y. (2025). Classification of Painting Image Styles Based on Hybrid Vision Transformer–ResNet50 Framework. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 1-7). IEEE. https://doi.org/10.1109/ic3it66137.2025.11341636

[17] Wei, L., Li, Q., & Wang, D. (2026). Multitask Art Painting Categorization by Convolution and Transformer Hybrid Architecture. Knowledge Organization, 53(2). https://doi.org/10.31083/ko48360

[18] Wang, H., Huang, Y., (2025). Using deep learning algorithms to identify diverse types of art designs. International Journal of Information and Communication Technology, 26(24), 15-28. https://doi.org/10.1504/ijict.2025.10071894

[19] Zhang, Y., Lv, F., (2025). Enhancing digital art style recognition via a hybrid vision transformer and lightweight CNN with attention mechanisms. International Journal of Innovative Computing and Applications, 15(4), 236-245. https://doi.org/10.1504/ijica.2025.10073624

[20] Wang, Z., & Song, H. (2025). A fusion model for artwork identification based on convolutional neural networks and transformers. arXiv preprint arXiv:2502.18083. https://doi.org/10.48550/arXiv.2502.18083

[21] Parres, D., Paredes, R., (2023). Fine-tuning vision encoder-decoder transformers for handwriting text recognition on historical documents. In: Document Analysis and Recognition – ICDAR 2023. Lecture Notes in Computer Science, vol. 14190. Springer, Cham, 253-268. https://doi.org/10.1007/978-3-031-41685-9_16

[22] Roald, M., Birkenes, M.B., Johnsen, L.B., (2024). Visual navigation of digital libraries: Retrieval and classification of images in the National Library of Norway's digitised book collection. In: Proceedings of the 5th Conference on Computational Humanities Research (CHR24), Aarhus, Denmark, December 04-06, 2024. https://doi.org/10.48550/arXiv.2410.14969%20Focus%20to%20learn%20more

[23] Sivan, R., Pati, P.B. and Kesiman, M.W.A., (2025). Image quality determination of palm leaf heritage documents using integrated discrete cosine transform features with vision transformer. International Journal on Document Analysis and Recognition (IJDAR), 28(1), 41-57. https://doi.org/10.1007/s10032-024-00490-x

[24] Vidal-Gorène, C., Decours-Perez, A., (2024). Detecting and deciphering damaged medieval Armenian inscriptions using YOLO and vision transformers. In: Document Analysis and Recognition – ICDAR 2024 Workshops, Lecture Notes in Computer Science, 14941, 152-168. https://doi.org/10.1007/978-3-031-70642-4_2

[25] Schaerf, L., Popovici, C., Postma, E., (2024). Art authentication with vision transformers. Neural Computing and Applications, 36(20), 11849-11858. https://doi.org/10.1007/s00521-023-08864-8

[26] Capriotti, S., Devoto, A., Scardapane, S., Mignardi, S., & Medeghini, L. (2025). Interpretable classification of Levantine ceramic thin sections via neural networks. Machine Learning: Science and Technology, 6(2), 025076. https://doi.org/10.1088/2632-2153/ade6c4

[27] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., (2021). Training data-efficient image transformers & distillation through attention. In: Proceedings of the 38th International Conference on Machine Learning (ICML), 10347-10357. https://proceedings.mlr.press/v139/touvron21a

[28] Hassani, A., Walton, S., Shah, N., Abuduweili, A., Li, J., Shi, H., (2021). Escaping the Big Data Paradigm with Compact Transformers. arXiv:2104.05704. https://doi.org/10.48550/arXiv.2104.05704%20Focus%20to%20learn%20more

[29] Yu, L., (2026). Visual perception based deep learning transformers for classifying paintings and photographs through feature extraction. Scientific Reports (under publication). https://doi.org/10.1038/s41598-026-36298-4

[30] Hong, S., Wu, J., Zhu, L. and Chen, W., (2024). Brain tumor classification in VIT-B/16 based on relative position encoding and residual MLP. Plos one, 19(7), p.e0298102. https://doi.org/10.1371/journal.pone.0298102

[31] Shahi, M., Rajabi, R., Masoumzadeh, F., (2024). CNN-based classification of Persian miniature paintings from five renowned schools. Digital Applications in Archaeology and Cultural Heritage, 36, e00397. https://doi.org/10.1016/j.daach.2024.e00397

[32] Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412. https://doi.org/10.48550/arXiv.1710.09412

[33] Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101. https://doi.org/10.48550/arXiv.1711.05101

[34] Tang, M., Ting, K.C. and Rashidi, N.H., (2024). DenseNet201-based waste material classification using transfer learning approach. Applied Mathematics and Computational Intelligence (AMCI), 13(2), 113-120. https://doi.org/10.58915/amci.v13i2.555

[35] Vulli, A., Srinivasu, P.N., Sashank, M.S.K., Shafi, J., Choi, J. and Ijaz, M.F., (2022). Fine-tuned DenseNet-169 for breast cancer metastasis prediction using FastAI and 1-cycle policy. Sensors, 22(8), 2971-2988. https://doi.org/10.3390/s22082988

[36] Khavani, S. and Marasi, M., (2025). The History of Iranian Painting as a Function of the History of Persian Literature: An Analysis of Realism in the Works of Kamal al-Din Behzad Based on the Shifts in Literary Discourse. Rahpooye Honar-Ha-Ye Tajassomi, 8(2), 45-58. https://rahpooye.soore.ac.ir/article_728299.html

[37] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 618-626. https://doi.org/10.1109/iccv.2017.74