International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-11

https://doi.org/10.70917/ijcisim-2026-2821
Article

Multimodal Representation Learning and Visual Analytics for Modeling the Formal Visual Language of Contemporary Inner Mongolian Painting

Shuzhen Yan [1,*]

[1] School of Fine Arts and Design, Jining Normal University, Ulanqab, Inner Mongolia, 830000, China
* Correspondence author: 18648058621@163.com

Abstract: This paper presents a computer vision-driven multimodal representation learning framework for modeling the formal visual language of contemporary Inner Mongolian painting. A corpus of approximately 1,200 digitized paintings is curated together with structured metadata collected from museums, regional galleries, exhibition catalogues, and artist archives. To move beyond descriptive digital archiving, we formulate artwork analysis as a joint image–metadata learning problem. Specifically, painting images are encoded by a deep visual backbone, while artwork-level metadata, including artist, year, genre, subject matter, medium, and descriptive tags, are encoded into semantic representations. The two modalities are projected into a shared visual-semantic latent space through contrastive alignment and further integrated by a metadata-aware fusion module. The resulting joint representations are evaluated on style classification, semantic retrieval, unsupervised clustering, and visual analytics. Compared with an image-only CNN baseline, the proposed multimodal framework improves style classification accuracy from 78% to 88%, semantic consistency from 65% to 80%, feature extraction accuracy from 70% to 85%, and Precision@10 from 70% to 82%. Qualitative analyses based on confusion matrices, embedding visualization, radar plots, and theme co-occurrence graphs further show that the learned representation better separates stylistic categories while preserving culturally meaningful motifs and thematic relations. The study demonstrates that multimodal representation learning can provide an interpretable computational pipeline for regional artwork analysis and offers a data-driven bridge between computer vision and digital humanities.

Keywords: Digital Humanities; Art Style Analysis; Multimodal Embedding; Visual Analytics

1. Aims and Background

Modeling the formal visual language of paintings is a challenging computer vision problem because artistic style is not determined by a single visual cue, but by the joint organization of color, texture, brushstroke, composition, motif recurrence, and cultural semantics. Traditional art-historical analysis provides rich interpretation [1-2] but it is difficult to scale to large image collections and cannot easily support quantitative retrieval or representation-level comparison.

Existing studies on computational artwork analysis have achieved considerable progress in tasks such as artist identification, style recognition, and content-based image retrieval. However, many of these approaches still handle visual information and contextual metadata as two relatively independent sources. Such a separation is particularly limiting for regional art collections, including contemporary Inner Mongolian painting, where visual motifs, cultural symbols, titles, themes, and historical context jointly shape the meaning of an artwork. For this reason, we treat the computational study of Inner Mongolian painting as a multimodal representation learning problem. In this setting, painting images and structured metadata are encoded together so that the resulting representations can support style classification, semantic retrieval, clustering, and visual analytics [3].

The development of computational artwork analysis has moved from hand-crafted visual features toward deep learning-based representations. Earlier content-based retrieval systems usually described paintings through low-level visual cues, including color histograms, texture features, edge distributions, and shape descriptors. These features were then used to estimate the visual similarity between artworks [4]. Related studies also explored more specialized visual properties of paintings. For example, brushstroke rhythm and color usage have been used to distinguish the works of different artists. Li et al. showed that the rhythm of brushstrokes could help separate Vincent van Gogh's paintings from those of other painters [4]. Later, the availability of larger annotated art datasets, such as Painting, as well as benchmark tasks such as the Rijksmuseum Challenge for multi-category art classification, encouraged the development of more advanced methods [5].

With the rapid growth of deep learning, convolutional neural networks became a dominant tool for artwork classification and recognition. Researchers began to fine-tune CNN models on art image datasets and reported clear improvements in recognizing artistic periods, genres, and styles [6]. Bar et al. demonstrated that deep CNN features could be used effectively for artistic style classification [7]. Wilber et al. further introduced the BAM dataset, which extended visual recognition beyond ordinary photographic images and covered multi-label artwork-related tasks such as style, genre, and medium prediction. Although these methods significantly improved classification accuracy, most of them still relied mainly on visual features or treated metadata as an external indexing tool. Image-only models can group paintings according to visual appearance, but their results are not always easy to interpret from an art-historical perspective. Conversely, metadata-based organization by artist, school, or period may overlook the actual visual patterns, motifs, and compositional characteristics of the artworks. As Manovich has argued, traditional metadata categories often fail to capture the visual structures that shape how artworks look and how they are perceived. This limitation points to the need for computational approaches that can connect visual representation with the broader contextual information attached to artworks [7-9].

In response to this gap, we embed digital humanities questions within a reproducible computer-vision framework. In our approach, visualization is not treated merely as the final output of analysis. Instead, multimodal representation learning serves as the central computational mechanism, while visual analytics functions as an interpretive layer built upon the learned embeddings. Each painting is therefore represented through both its visual appearance and its structured contextual information, including artist, production year or period, subject matter, medium, genre, and cultural tags. By aligning these heterogeneous sources of information within a shared visual-semantic space, the proposed framework can support both machine-learning-oriented tasks, such as classification and retrieval, and humanities-oriented analysis, such as motif discovery, thematic comparison, and co-occurrence analysis. In this way, regional painting style can be investigated as a measurable representation-learning problem without removing the cultural context required for interpretation [8].

This perspective is closely related to recent developments in digital humanities. Scholars have proposed the idea of "distant viewing," which is comparable to distant reading in literary studies, as a way to examine large image collections computationally and identify patterns that may not be immediately visible through close observation alone. Arnold and Tilton's work on distant viewing demonstrates how computational methods can reveal broad visual tendencies across image corpora and thereby support art-historical analysis at scale. For artistic images, however, meaningful computational analysis should not be limited to pixel-level patterns. It also needs to consider contextual information such as titles, themes, artist background, cultural references, and symbolic meaning.

Recent studies have increasingly emphasized this integration of visual and semantic information. Wevers and Smits showed that neural networks can categorize historical images effectively when trained with expert-labeled examples, reflecting a broader "visual digital turn" in humanities research. In computational art history, researchers have also begun to introduce metadata and knowledge-graph structures to improve both recognition performance and interpretability. For instance, Messina et al. proposed a content-based art recommendation system that combines painting metadata with deep visual features, resulting in recommendations that are more consistent with human preferences and easier to interpret. Garcia and Oramas introduced the SemArt dataset and multimodal retrieval methods for connecting fine-art images with textual descriptions, making it possible to retrieve paintings according to semantic or narrative themes [9]. These studies show a clear methodological trend: when artworks are embedded in a multimodal feature space that combines visual and semantic information, retrieval and clustering can move beyond surface-level visual similarity toward more meaningful conceptual similarity.

This direction is especially important for culturally specific art collections. In the case of Inner Mongolia, existing computational studies on visual heritage, such as rock art and intangible cultural symbols, have already begun to use deep learning to extract visual elements associated with cultural identity [10-12]. Nevertheless, contemporary Inner Mongolian painting has received much less attention

from computational methods. This art form often combines traditional Mongolian motifs with modern artistic techniques, making it difficult to analyze through either visual features or metadata alone. A digital humanities approach is therefore needed to model the formal language of this regional art while also preserving the cultural information carried by titles, stories, symbols, and themes [13-14]. Such an approach can help connect visual complexity with cultural interpretation and provide a more systematic basis for studying contemporary Inner Mongolian painting.

The contributions of this paper are threefold. First, we construct a multimodal artwork corpus for contemporary Inner Mongolian painting. The corpus links digitized painting images with structured metadata, including artist, year, genre, subject matter, medium, and descriptive tags. Second, we propose a computer vision-based multimodal representation learning framework. The framework integrates a deep visual encoder, a metadata encoder, cross-modal contrastive alignment, and metadata-aware feature fusion to produce joint visual-semantic embeddings for artworks. Third, we evaluate the learned representations through several downstream tasks, including style classification, semantic retrieval, unsupervised clustering, and visual analytics. The experimental results show that multimodal learning not only improves quantitative performance, but also enhances interpretability for regional art analysis.

2. Experimental Method

The proposed framework consists of five computational stages: multimodal corpus construction, modality-specific encoding, cross-modal contrastive alignment, metadata-aware feature fusion, and downstream visual knowledge discovery. Given an artwork image ($\boldsymbol{x}_i$) and its corresponding metadata record ($\boldsymbol{m}_i$), the visual branch extracts deep image features from the painting, while the metadata branch encodes artist-, year-, genre-, subject-, medium-, and tag-level information into a semantic representation. The two modality-specific representations are projected into a shared visual-semantic latent space and optimized by a contrastive objective. The aligned embeddings are then fused into a joint representation for style classification, semantic retrieval, unsupervised clustering, and interactive visual analytics.

A core component of the framework is a joint embedding model that aligns visual and semantic information. For each artwork (i), let $\boldsymbol{x}_i$ denote the image and $\boldsymbol{m}_i$ denote the associated metadata record. We encode the image with a visual encoder $\boldsymbol{f}_\theta$ to obtain a feature vector $\boldsymbol{v}_i = \boldsymbol{f}_\theta(\boldsymbol{x}_i)$, and we encode the metadata with a semantic encoder $\boldsymbol{g}_\phi$ to obtain $\boldsymbol{t}_i = \boldsymbol{g}_\phi(\boldsymbol{m}_i)$. Two projection heads map the modality-specific features into a shared (d)-dimensional latent space: $\boldsymbol{z}_i^{img} = \text{norm}(\boldsymbol{W}_v \boldsymbol{v}_i)$ and $\boldsymbol{z}_i^{meta} = \text{norm}(\boldsymbol{W}_t \boldsymbol{t}_i)$. Training minimizes a symmetric contrastive objective that pulls matched image–metadata pairs together while pushing unmatched pairs apart within each mini-batch. After training, the paired embeddings are fused into a metadata-aware representation that is used for style classification, semantic retrieval, clustering, and visualization.

$$\boldsymbol{v}_i = \boldsymbol{f}_{\text{img}}(\boldsymbol{I}_i; \boldsymbol{\theta}_{\text{img}}) \in \mathbb{R}^{d_v} \tag{1}$$

where $\boldsymbol{f}_{\text{img}}$ is the image feature extraction network (e.g., a ResNet or ViT model) parameterized by $\boldsymbol{\theta}_{\text{img}}$. Typically, $\boldsymbol{f}_{\text{img}}$ includes convolutional layers followed by a pooling layer that produces a compact representation of the painting's visual content (such as brushstroke textures, color distributions, and compositional structures). Likewise, we encode the metadata $\boldsymbol{M}_i$ (which may be a text string or a set of tags) into a $\boldsymbol{d}_m$-dimensional semantic feature vector:

$$\boldsymbol{u}_i = \boldsymbol{f}_{\text{meta}}(\boldsymbol{M}_i; \boldsymbol{\theta}_{\text{meta}}) \in \mathbb{R}^{d_m}, \tag{2}$$

With $\boldsymbol{f}_{\text{meta}}$ representing the metadata encoder network (for example, a transformer-based language model or a simpler bag-of-words embedding model) with parameters $\boldsymbol{\theta}_{\text{meta}}$. This function converts textual or categorical information about the painting into a numerical vector capturing its semantic attributes (such as themes, motifs, or stylistic keywords present in $\boldsymbol{M}_i$).

Because the image and metadata features originate from different modalities, we project them into a shared embedding space of dimension d to make them directly comparable. We employ learned linear projections (or small fully-connected networks) to map $\boldsymbol{v}_i$ and $\boldsymbol{u}_i$ into the same d-dimensional space. Let $\boldsymbol{W}_v \in \mathbb{R}^{d \times d_v}$ and $\boldsymbol{W}_u \in \mathbb{R}^{d \times d_m}$ be the projection matrices for image and metadata features,

respectively. We obtain:

$$z_i^I = W_v v_i \in \eth^d, z_i^M = W_u u_i \in \eth^d, \tag{3}$$

And in practice we apply ℓ_2 normalization so that z_i^I and z_i^M lie on the unit hypersphere. The vectors z_i^I and z_i^M can be interpreted as the visual-semantic embeddings of the i-th painting: z_i^I is the image-derived embedding and z_i^M is the metadata-derived embedding. By design, these embeddings live in the same feature space, which enables measuring their similarity directly (for instance, via dot product or cosine similarity). We define the similarity between an image I_i and a metadata description M_j as the dot product of their normalized embeddings.

$$L_{\text{I2M}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\text{sim}\left(I_i, M_i\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}\left(I_i, M_j\right)/\tau\right)}, \tag{4}$$

$$L_{\text{M2I}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\text{sim}\left(M_i, I_i\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}\left(M_i, I_j\right)/\tau\right)}, \tag{5}$$

$$L_{\text{contrastive}} = \frac{1}{2}\left(L_{\text{I2M}} + L_{\text{M2I}}\right), \tag{6}$$

Which we minimize with respect to the encoder parameters ($\theta_{\text{img}}, \theta_{\text{meta}}$) and the projection matrices (W_v, W_u). By minimizing $L_{\text{contrastive}}$, the model learns to align z_i^I and z_i^M for each painting i, effectively embedding images and their semantic metadata closely together.

3. Results and Discussion

3.1. Quantitative evaluation of multimodal representation quality

To evaluate the quality of multimodal representations, we constructed a curated research corpus of contemporary Inner Mongolian paintings. The corpus brings together digitized artworks collected from museums, regional galleries, exhibition catalogues, and artist archives, and organizes them into a unified repository using unique artwork identifiers. For each painting, the dataset records a normalized image together with structured metadata, including artist, year, genre, subject matter, medium, and descriptive tags. In the preprocessing stage, the images were processed through acquisition, color calibration, cropping, resolution normalization, and format conversion, and were stored in both archival and web-compatible versions. Through this design, the corpus serves not only as a digital archive but also as a dataset suitable for machine-learning tasks such as representation learning, retrieval, clustering, and visual analytics in the relatively under-explored field of regional Inner Mongolian art.

Data scale and contents. The final corpus contains approximately 1,200 digitized paintings dating from the late twentieth century to the present. It covers a range of genres and stylistic categories commonly found in the work of contemporary Inner Mongolian artists. Figure 1 presents the distribution of the main thematic categories in the dataset. Grassland Landscapes account for the largest proportion, representing around 30% of the works, followed by scenes of Nomadic Life, which make up about 25%. These two categories reflect the central role of pastoral and nomadic culture in Inner Mongolian painting. Motifs related to Horse Culture appear in approximately 15% of the paintings, often in connection with herding life and grassland imagery. Folklore & Myth also accounts for roughly 15% of the dataset, and many of these works overlap thematically with pastoral or traditional cultural scenes. In comparison, Urban Scenes are less frequent, comprising under 10% of the corpus. This suggests that contemporary Inner Mongolian painters continue to place stronger emphasis on rural life, cultural memory, and regional heritage than on modern urban experience. The remaining Other category, approximately 5%, includes experimental abstract works and paintings that do not fit clearly into the dominant thematic groups.

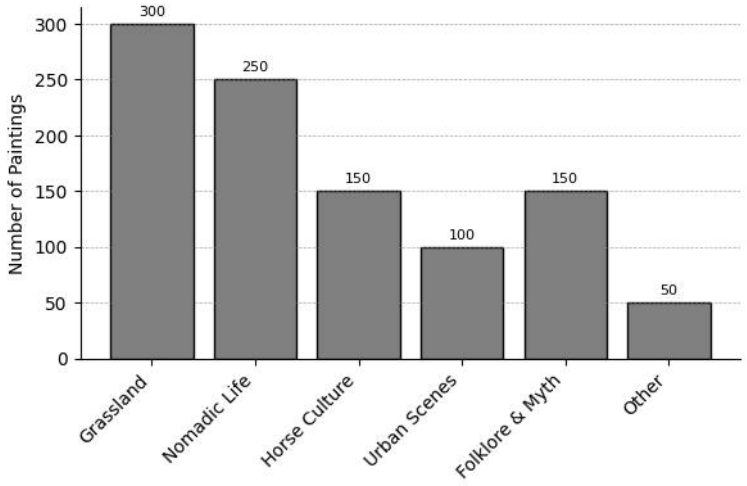

Figure 1. Distribution of painting themes in the constructed Inner Mongolian art database (number of paintings per theme).

This distribution indicates that the corpus captures a meaningful cross-section of the major visual and cultural concerns in contemporary Inner Mongolian painting. The dataset is strongly oriented toward nature, animals, pastoral life, and cultural tradition, while works focused on urbanization or purely abstract expression are relatively limited. As a result, the archive provides a useful basis for computational analysis, enabling systematic comparison of artworks that were previously scattered across different institutions and sources. Overall, this section describes the construction of a novel Inner Mongolian painting dataset, supported by standardized digitization procedures and structured metadata that help represent the formal and cultural language of each work.

3.2. Experimental comparison of proposed method vs. Baselines

Visual style classification. We first tested the learned representations on a four-class style classification task, covering realist, expressionist, folk-inspired, and abstract paintings. The image-only baseline, based on a ResNet-50 backbone, achieved a test accuracy of 78%. By comparison, the proposed multimodal framework achieved 88% accuracy under the same evaluation setting, corresponding to a 10 percentage-point absolute improvement. This result shows that incorporating both visual information and contextual metadata can produce more discriminative representations than relying on image features alone, especially for the analysis of region-specific artistic styles.

As shown in Figure 2, the classification accuracy of the baseline and the proposed method is compared in the left-most cluster of bars. The baseline reaches 78%, whereas the proposed method reaches 88%. This improvement suggests that formal metadata and cultural contextual information provide useful complementary cues for distinguishing Inner Mongolian art styles. Although the CNN baseline is able to capture general visual patterns, it may have difficulty identifying subtle differences between styles that share similar textures, colors, or compositions. In contrast, the proposed method benefits from additional signals such as brushstroke characteristics, palette information, and culturally meaningful motifs, which help reduce ambiguity in style recognition.

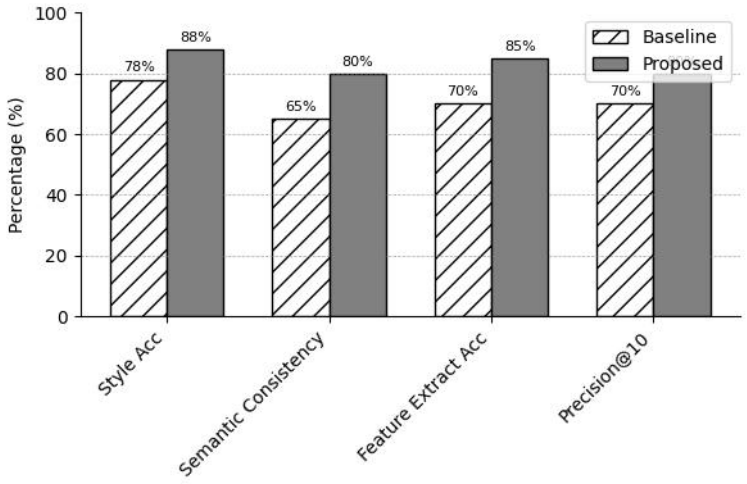

Figure 2. Performance comparison of the proposed method vs. a baseline CNN.

To further examine the classification results, we report the confusion matrices for style prediction. In these matrices, darker diagonal cells indicate higher correct classification rates. In the baseline confusion matrix shown in Figure 3a, most samples are still correctly classified, such as 80% of Realist paintings being predicted as Realism. However, several noticeable confusions remain. For example, Abstract paintings are sometimes misclassified as Expressionist works: only 65% of Abstract paintings are correctly identified, while 20% are incorrectly assigned to Expressionism. A similar pattern can be observed for Folk-traditional works, some of which are confused with Expressionist or Abstract styles. These off-diagonal errors, often reaching 15–20%, indicate that the image-only baseline struggles to separate styles with overlapping visual characteristics.

The proposed method shows a clearer diagonal pattern in the confusion matrix in Figure 3b. Specifically, it correctly classifies 90% of Realist paintings, 85% of Expressionist paintings, 80% of Abstract paintings, and 87% of Folk paintings. At the same time, the off-diagonal error rates are reduced. For instance, the confusion between Abstract and Expressionist works decreases from 20% in the baseline to approximately 10% in the proposed method. This improvement indicates that the representation learned by the proposed framework is more effective in distinguishing visually or historically related styles. These findings are consistent with recent studies on automated art classification, where carefully designed CNN-based models have achieved high accuracy, often exceeding 90%, while also showing that some art styles, such as Abstract and Surrealism, may overlap in feature space. In our case, the proposed method appears to reduce such overlap by combining visual texture with semantic and cultural cues, including motifs and metadata, rather than relying only on pixel-level image features.

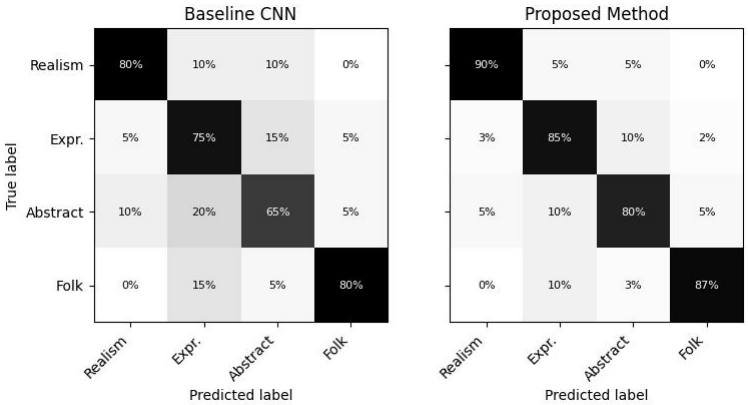

Figure 3. Confusion matrices for style classification: (a) Baseline CNN, (b) Proposed method.

3.3. Semantic consistency

We further evaluate whether the learned representation preserves semantic correspondence between visual content and metadata. A semantically reliable model should retrieve paintings with related concepts, such as horses, grassland scenes, nomadic life, folklore, or religious imagery, when queried by

either visual similarity or metadata-level attributes. Under this evaluation, the image-only baseline achieves 65% semantic consistency, whereas the proposed multimodal method reaches 80%. The improvement indicates that metadata-aware fusion enhances not only visual discrimination but also concept-level retrieval and thematic coherence. This result is particularly important for digital humanities applications, where researchers often need to search, compare, and interpret artworks according to culturally meaningful motifs rather than only visual appearance.

3.4. Clustering of stylistically similar works

To further compare the internal representations learned by the two approaches, we performed unsupervised clustering experiments. We used t-distributed Stochastic Neighbor Embedding (t-SNE) to visualize the painting embeddings produced by each method. Figure 4 shows the t-SNE plot for the proposed method's feature space, where each point represents a painting and points are plotted with different markers for the four major style categories

As shown in Fig 4, Clusters of points illustrate that paintings naturally group by style under the proposed representation. We see that the proposed method yields well-separated clusters: for example, the Realist paintings (circles) cluster together in one region, clearly apart from the Abstract paintings (triangles) which form another cluster. Expressionist works (squares) also group in their own space, though with a slight overlap toward Realist clusters (likely reflecting some intermediate characteristics), and Folk style paintings (diamonds) cluster tightly, indicating consistent feature capture for those traditional works. This visualization suggests that the features learned by our method carry meaningful high-level information that groups similar artworks together – the model has learned a latent space respecting art-historical categories.

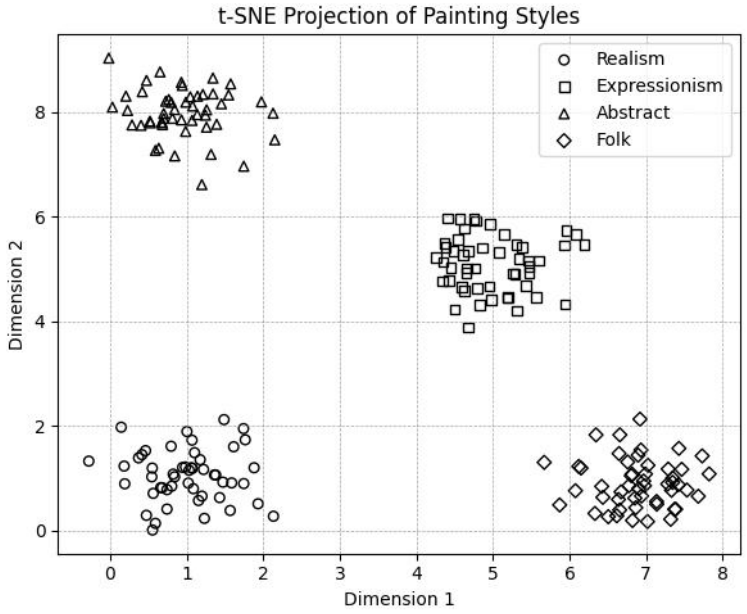

Figure 4. *t*-SNE projection of paintings based on learned features (proposed method). Each point is a painting embedded in 2D; distinct markers denote known style categories (Realism ○, Expressionism □, Abstract △, Folk ◇)

3.5. Experimental Comparison of Proposed Method vs. Baselines

3.5.1. Comparative stylistic profiles (radar chart)

To understand differences in formal language across art styles, we plotted a radar chart (also known as a spider chart) comparing the average feature values for several style categories. As shown in Fig 5. Axes represent normalized feature scores (0–1) for Color usage, Brushstroke boldness, Composition complexity, Motif richness, and Texture detail. Each style is represented by a polygon connecting its scores on the five axes. Several clear patterns emerge from Figure 5. The Expressionist style (dashed line with square markers) scores highest on Color (around 0.9) and Brushstroke (0.8), reflecting that these paintings use very vibrant palettes and energetic, bold brushwork. However, Expressionism has a lower Motif score (~0.3), indicating it tends to emphasize form and color over narrative content – in other

words, expressive works are more about how something is painted than what is painted. In contrast, the Folk-inspired style (dash-dot line with diamond markers) excels in Motif richness (0.9), meaning folk style paintings incorporate many symbolic elements and narrative motifs (e.g. traditional patterns, mythological figures), but they use comparatively subdued Color (0.5) and simpler Brushstrokes (0.4), often favoring flat decorative application over painterly texture. Realist works (solid line with circle markers) show a balanced profile: moderate color intensity (0.6), moderate brushstroke (0.5), but the highest Composition complexity (0.8) – consistent with Realist artists carefully constructing perspective and detailed arrangements in their paintings. Realism also has relatively high Texture (0.7), as these works pay attention to fine details and material effects. Finally, Abstract style paintings (dotted line with triangle markers) unsurprisingly score high on Texture (0.8, the highest among styles, as many abstract works experiment with thick paint or mixed media textures) and fairly high on Color (0.7), but low on Motif (0.2, since abstraction minimizes recognizable subjects) and lower on Composition (0.4, often deliberately breaking formal compositional rules).

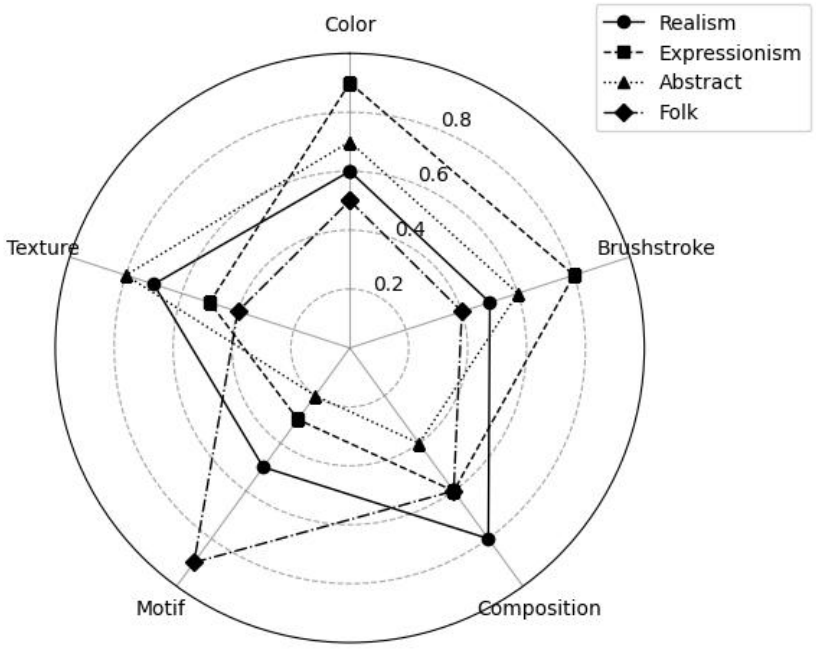

Figure 5. Radar chart comparing five formal features across four painting style groups (Realism, Expressionism, Abstract, Folk).

3.5.2. Semantic network of themes (co-occurrence graph)

In addition to formal features, we explored the relationships between content themes in the paintings using network analysis. To examine thematic relationships within the corpus, we built a co-occurrence network in which each node denotes a major theme, subject, or motif, while edges indicate that two themes appear together in the same painting with relatively high frequency. As shown in Figure 6, traditional cultural themes are closely interlinked, whereas the Urban theme remains almost separate from the rest of the network, connected only by a very weak edge. The network therefore provides an overall view of how different visual themes are associated with one another in contemporary Inner Mongolian painting.

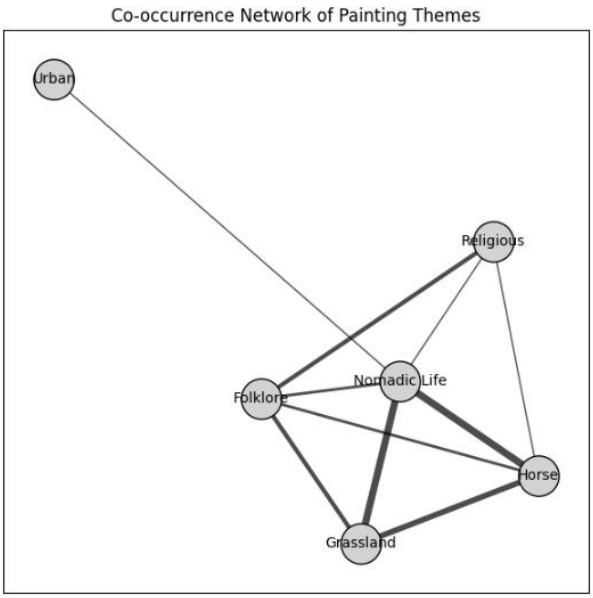

Figure 6. Co-occurrence network of major painting themes. Nodes are theme keywords (Nomadic Life, Horse, Grassland, Folklore, Religious, Urban)

Several patterns are evident from the network structure. Nomadic Life occupies a central position in the main thematic cluster and is strongly connected with both Horse and Grassland, as reflected by the thicker edges. This pattern provides quantitative support for a long-standing cultural observation: paintings in this corpus often present pastoral life, horses, and grassland scenery as an integrated visual subject rather than as isolated elements. The strongest connections appear between Nomadic–Horse and Nomadic–Grassland, indicating that scenes of herders, horses, and open steppe landscapes constitute one of the most common representational forms in the dataset. The direct link between Horse and Grassland further shows that horses are frequently depicted within natural grassland settings, even when human figures are absent.

The network also reveals more specific thematic combinations. The moderate edge between Nomadic Life and Folklore suggests that some works combine everyday pastoral scenes with legendary, symbolic, or folk-cultural narratives. Folklore, in turn, is strongly associated with Religious imagery, as indicated by the relatively thick edge between these two nodes. This implies that mythological or legendary subjects in the corpus are often accompanied by spiritual or Buddhist visual elements. For example, a painting may portray a heroic or legendary figure while also incorporating religious symbols or iconographic references. By contrast, the direct connection between Religious imagery and Nomadic Life is weak. This suggests that religious works, such as Buddhist or thangka-related paintings, generally form a relatively independent category and only occasionally overlap with scenes of everyday pastoral life, unless mediated through folklore or mythological subject matter.

The Urban node shows a notably different pattern. It is nearly isolated, with only a single weak connection to Nomadic Life. This indicates that only a small number of paintings attempt to place modern urban elements alongside nomadic or pastoral culture. In most cases, urban scenes appear separately rather than being integrated with traditional themes. This result is consistent with the earlier thematic distribution of the dataset, where urban subjects account for only a small proportion of the corpus. The marginal position of Urban in the co-occurrence network therefore makes visible a broader thematic separation between contemporary urban experience and the dominant pastoral, cultural, and historical concerns of Inner Mongolian painting.

Overall, the experimental results demonstrate that multimodal representation learning offers a stronger computational foundation for the study of contemporary Inner Mongolian painting than image-only modeling. Compared with the baseline approach, the proposed framework improves style classification, strengthens semantic coherence, and generates embedding structures that correspond more closely to stylistic and thematic distinctions in the corpus. Figure 7 shows the muti-modal computational framework for image-metadata representation learning and visual knowledge discovery.

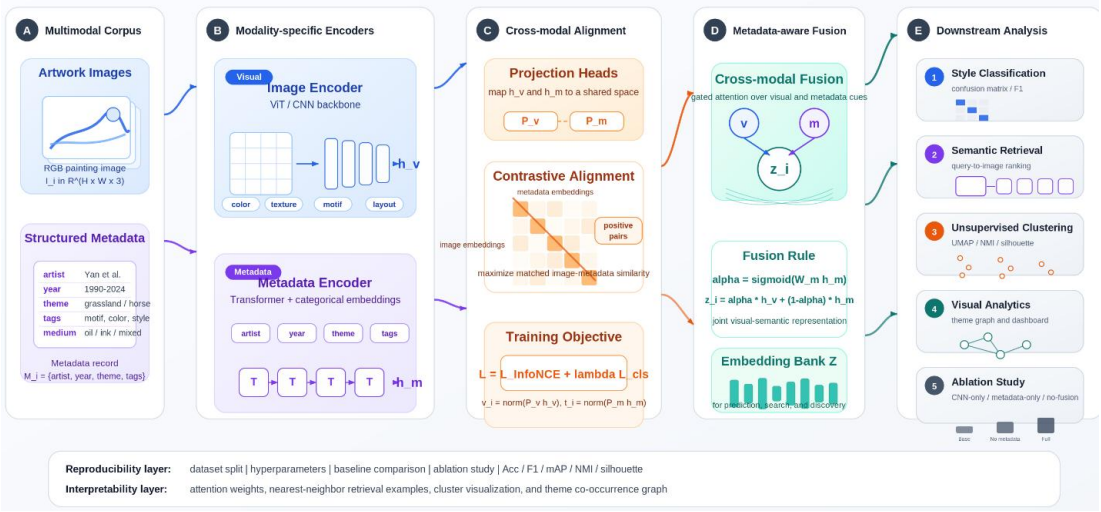

Figure 7. Muti-modal computational framework for image-metadata representation learning and visual knowledge discovery.

Moreover, visual analytic methods, including embedding visualization, radar charts, and theme co-occurrence networks, make the learned representations more interpretable from an art-historical perspective. These results suggest that computer vision and multimodal learning can support digital humanities research not only by organizing visual archives, but also by enabling scalable, reproducible, and interpretable discovery of visual and cultural knowledge.

Acknowledgement

Project Name: Research on the Expression of Contemporary Inner Mongolia Painting - Archive through the website, Project number: jsbsjj2420, Project leader: Yan Shuzhen, College (Department): School of Fine Arts and Design, Jining Normal University.

References

1. T. Arnold and L. Tilton, “Distant viewing: analyzing large visual corpora,” Digital Scholarship in the Humanities, vol. 34, no. S1, pp. i3–i16, 2019.
2. L. Manovich, “Data science and digital art history,” International Journal for Digital Art History, no. 1, pp. 12–35, 2015.
3. M. Wevers and T. Smits, “The visual digital turn: using neural networks to study historical images,” Digital Scholarship in the Humanities, vol. 35, no. 1, pp. 194–207, 2020.
4. L. Shamir, T. Macura, N. Orlov, D. M. Eckley, and I. G. Goldberg, “Impressionism, expressionism, surrealism: automated recognition of painters and schools of art,” ACM Transactions on Applied Perception, vol. 7, no. 2, pp. 1–17, 2010.
5. R. S. Arora and A. Elgammal, “Towards automated classification of fine-art painting style: a comparative study,” in Proceedings of the 21st International Conference on Pattern Recognition (ICPR), 2012, pp. 3541–3544.
6. F. S. Khan, S. Beigpour, J. van de Weijer, and M. Felsberg, “Painting-91: a large scale database for computational painting classification,” Machine Vision and Applications, vol. 25, no. 6, pp. 1385–1397, 2014.
7. Y. Bar, N. Levy, and L. Wolf, “Classification of artistic styles using binarized features derived from a deep neural network,” in Computer Vision – ECCV 2014 Workshops, Zurich, Switzerland: Springer, 2015, pp. 71–84.
8. M. J. Wilber, C. Fang, H. Jin, A. Hertzmann, J. Collomosse, and S. Belongie, “BAM! The Behance Artistic Media dataset for recognition beyond photography,” in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1202–1211.

9. E. Cetinic, T. Lipic, and S. Grgic, “Fine-tuning convolutional neural networks for fine art classification,” Expert Systems with Applications, vol. 114, pp. 107–118, 2018.
10. E. Gultepe, T. E. Conturo, and M. Makrehchi, “Predicting and grouping digitized paintings by style using unsupervised feature learning,” Journal of Cultural Heritage, vol. 31, pp. 13–23, 2018.
11. P. Messina, V. Dominguez, D. Parra, C. Trattner, and A. Soto, “Content-based artwork recommendation: integrating painting metadata with neural and manually-engineered visual features,” User Modeling and User-Adapted Interaction, vol. 29, no. 2, pp. 251–290, 2019.
12. N. Garcia and J. V. Oramas, “How to read paintings: semantic art understanding with multi-modal retrieval,” in Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, pp. 0–0.
13. X. Bu and M. Jiang, “Extraction of intangible cultural heritage visual elements by deep learning,” Computational Intelligence and Neuroscience, vol. 2022, Art. ID 8567895, 2022.
14. X. Du and Y. Cai, “Design of Chinese painting style classification model based on multi-layer aggregation CNN,” PeerJ Computer Science, vol. 10, article e2303, 2024.