Knowledge Organization

Knowl. Organ. **2026**; 53(2): 48360
https://doi.org/10.31083/KO48360

Article

Multitask Art Painting Categorization by Convolution and Transformer Hybrid Architecture

Liangyu Wei[1,*], Qianheng Li[1], Dandan Wang[1]

[1]Academy of Fine Arts, Shandong Normal University, 250061 Jinan, Shandong, China
*Correspondence: 2025120059@stu.sdnu.edu.cn (Liangyu Wei)
Academic Editor: Natália Tognoli
Submitted: 18 November 2025 Revised: 22 February 2026 Accepted: 27 March 2026 Published: 23 April 2026

Liangyu Wei received her Bachelor of Arts degree from Jilin Normal University in 2017, and her Master of Arts degree from Shandong Normal University in 2020. Currently pursuing a Doctor degree at Shandong Normal University. Her research interests include art aesthetics, art education and art history. She has won multiple prominent awards: the first prize at the Annual Conference of Chinese Art History, the first prize at the Shandong Newcomers and New Works Exhibition. She has published several high-level papers and obtained two patents.

Qianheng Li received a Bachelor of Arts degree and a Master of Arts degree from Shandong Normal University in 2013 and 2017. The main research directions are art history and art practice. She has won many awards, including the Excellence Award in the National Youth Art Star Calligraphy and Painting Competition, the Guan Qunfang Cup Second Prize in the Second China Shandong Youth Calligraphy and Painting TV Competition. She was selected for the 2016 First Shandong (International) ContemporaryArt Exhibition.

Dandan Wang received the Master of Arts degree from Shandong Normal University in 2016. She has published several high-level papers and monograph. Her research interests include art aesthetics, art education and art history.

Abstract

Automatic categorization of fine art paintings across multiple semantic facets, such as artist, style, and genre, is fundamental for large-scale digital archiving, semantic indexing, and knowledge organization of cultural heritage collections. In this paper, we propose convolutional neural network (CNN)-Transformer Hybrid Attention model for art paintings categorization (CTHArt), a CNN-Transformer Hybrid Attention network for multitask art painting categorization. The model employs a dual-branch hybrid backbone that combines a CNN stream for fine-grained local texture modeling and a Transformer stream for global compositional and stylistic context learning. To further exploit inter-facet semantic dependencies, we introduce a Cross-Task Attention Head, which enables task-specific classifiers to exchange information through learnable cross-attention interactions. This design supports coordinated facet prediction consistent with knowledge organization principles. We evaluate the proposed framework on three benchmark datasets. Experimental results demonstrate that CTHArt consistently achieves state-of-the-art performance. The proposed approach provides an effective and scalable solution for artificial intelligence (AI)-assisted knowledge organization of art collections.

Keywords: Chinese traditional painting classification; artificial intelligence; convolutional neural network; Vision Transformer; cross attention

1. Introduction

The automatic categorization of fine art paintings (including artist attribution, style classification, and genre identification) plays a critical role in large-scale digital archiving, semantic indexing, and structured access to cultural heritage collections (Bianco et al., 2019; Specker et al., 2024). From a Knowledge Organization (KO) perspective, these labels function as complementary semantic facets that support systematic description, retrieval, and analysis of artworks (Hjørland, 2008). KO theory emphasizes that effective knowledge access depends on well-structured conceptual categorization and facet-based organization of domain entities (Giunchiglia et al., 2014; Giunchiglia and Bagchi, 2024). With the rapid growth of digital art repositories, artificial intelligence (AI)-driven methods are increasingly becoming essential components in knowledge organization ecosystems for scalable and consistent metadata construction (Bagchi, 2021a).

Painting analysis differs fundamentally from natural image understanding. Artistic representations often contain deliberate distortions, stylized textures, symbolic abstractions, and historically grounded visual conventions that challenge conventional computer vision assumptions (Zhao and Zhang, 2025; Xu et al., 2025). For example, artistic style encodes high-level aesthetic and historical characteristics such as Impressionism or Baroque rather than simple visual statistics. Artist attribution depends on subtle idiosyncratic cues such as brushstroke patterns and color layering strategies, while genre classification reflects thematic and compositional conventions. These semantic dimensions are not independent but form an interrelated knowledge.

Traditionally, painting categorization has been performed manually by experts, requiring specialized domain knowledge and extensive effort. Recent advances in deep neural networks (DNNs) have enabled automated approaches based on convolutional neural networks (CNNs) (González-Martín et al., 2024) and Vision Transformers (ViTs) (Dosovitskiy et al., 2020). However, existing methods typically treat artist, style, and genre prediction as isolated tasks and rely on single-architecture backbones, which introduces two important limitations (Gao et al., 2025). Firstly, artist, style, and genre are intrinsically linked (e.g., Van Gogh's post-impressionist works) in art paintings categorization task (as shown in Fig. 1), but single-task models fail to exploit shared and complementary semantic representations across multiple KO facets, reducing both learning efficiency and semantic consistency. Second, CNN-only or Transformer-only architectures provide incomplete feature modeling. CNNs are effective at capturing local visual structures such as textures and brushstrokes but have limited capacity for modeling long range compositional dependencies. Transformers capture global relationships through self-attention but usually require large-scale training data and higher computational cost, which is not always suitable for art datasets that are comparatively limited and fine-grained.

Recent hybrid CNN-Transformer architectures demonstrate that combining convolutional inductive bias with global attention modeling can yield more balanced representations. CNN layers provide stable local feature hierarchies, while Transformer modules model long-range structural and semantic dependencies. For example, 3M-Hybrid (Yang et al., 2025) integrates pretrained ViTs with CNNs for mural restoration, using frequency-based decomposition to handle scarce data and structural distortions. CDDFuse (Zhao et al., 2023) decomposes cross-modal features into correlated (low-frequency) and unique (high-frequency) components using CNN-Transformer dual branches, enhancing infrared-visible image fusion. These frameworks validate hybrid models' efficacy but remain unoptimized for multitask art analysis. Their fusion strategies lack explicit mechanisms to model task-specific interactions—e.g., how style features inform artist identification.

To address these limitations, and to better support AI-driven knowledge organization of art collections, we propose a CNN-Transformer Hybrid Attention model for art paintings categorization (CTHArt). The model is designed as a multitask, multibranch architecture that jointly predicts artist, style, and genre labels while modeling their semantic interactions. The hybrid backbone integrates a CNN branch for high-frequency local artistic details and a Transformer branch for global compositional and stylistic context. Multiple classification tasks share this backbone to learn unified semantic representations. On top of the shared features, we introduce a cross-task attention mechanism between task-specific decoder heads, enabling each task to selectively attend to informative features from the other tasks. This design implements flexible inter-facet semantic interaction rather than a fixed hierarchical dependency, aligning with KO principles of facet coordination and semantic linking. We evaluate the proposed framework on serval benchmark art datasets. Experimental results show that our model achieve state-of-the-art (SOTA) performance. Beyond performance gains, the proposed framework can also be viewed as an AI-assisted KO mechanism for cultural heritage data, enabling coordinated facet assignment and semantic enrichment in large-scale art knowledge bases (Bagchi, 2021b).

This work makes three key contributions:

• A CNN-Transformer hybrid backbone that jointly captures fine-grained artistic details and global stylistic semantics for multi-facet painting categorization.

• A cross-task attention mechanism is integrated into the multitask decoder, explicitly capturing inter-task dependencies through learnable feature interactions.

• Extensive experiments and ablation studies demonstrating the effectiveness of hybrid representation learning

Fig. 1. Some example art paintings (artist, style, and genre are intrinsically linked).

and cross-task semantic interaction for AI-driven art knowledge organization.

This paper is organized as follows: Section 2 reviews related works; Section 3 details our architecture; Section 4 evaluates performance against state-of-the-art baselines; Section 5 gives the conclusion and future work.

2. Related Works

This section should be clear and sufficiently detailed. Clearly outline the procedures, including a comprehensive explanation of data collection, analysis methods, and statistical approaches used. Specify the sources of any datasets, software, or tools utilized (e.g., database names, software versions, or vendors). If established methodologies are used, cite relevant references instead of providing extensive descriptions. Ensure clarity and precision in the presentation of methods.

2.1 Multitask Learning in Art Classification

Multitask learning has become a key paradigm in computational art analysis, exploiting shared visual features to improve classification across related tasks (Yang et al., 2022; Liu, 2024; Tian and Nan, 2022). Early efforts aggregated large art image collections and trained a single network to predict multiple attributes simultaneously. For example, Strezoski and Worring introduced OmniArt (Strezoski and Worring, 2018), a deep multi-task model with a shared representation for artistic data, and released a large-scale dataset of nearly half a million paintings with rich metadata. Their network was trained jointly on tasks such as artist, style, and material prediction, and was shown to outperform both hand-crafted features and single-task CNN baselines. Similarly, Bianco et al. (2019) proposed a Deep Multibranch CNN that processes different resolutions of the painting in parallel branches, solving artist, style, and genre classification in a unified network. This model was evaluated on the MultitaskPainting100k dataset (100K paintings, 1508 artists, 125 styles, 41 genres), demonstrating strong multi-task performance. These works highlight that jointly learning artist/style/genre helps the model leverage common representations. Recent reviews also note the growing use of multi-task deep learning and large art datasets for painting categorization (Ugail et al., 2023), underscoring that multi-task formulations (e.g., shared backbones or feature exchange) can capture inter-task correlations and improve overall accuracy.

2.2 CNN-Transformer Hybrid Models

Hybrid networks that combine convolutional layers and self-attention (Transformer) modules have gained traction for visual tasks (Arshad et al., 2024; Fang et al., 2022; Chen et al., 2024). These hybrids aim to marry the local detail encoding of CNNs with the global context modeling of Transformers. One representative example is the Convolutional vision Transformer (CvT), which integrates convolutional token embeddings and convolutional projec-

tions into a ViT-style model (Wu et al., 2021). By doing so, CvT inherits desirable invariances from CNNs (e.g., shift/scale invariance) while retaining the dynamic attention and long-range reasoning of Transformers. Empirically, CvT achieved state-of-the-art image classification performance on ImageNet with fewer parameters and Floating Point Operations (FLOPs) than comparable ViT models. More generally, Long (2024) survey the design space of CNN-Transformer hybrids and observe that these architectures can capture multi-scale features, "combining the local features extracted by CNNs with the global features learned by ViTs" to yield strong performance. This synergistic effect has been exploited in diverse domains: for instance, hybrid CNN-ViT models have set new benchmarks in medical and remote-sensing image analysis by capturing both fine-grained texture and global layout. Notably, Shah et al. (2024) demonstrate a practical gain: a three-branch hybrid (stacking CNN and ViT encoders) significantly outperforms a pure ViT on a multi-class disease classification task, confirming that fusing CNN and Transformer features improves discriminative power. These studies motivate our use of a dual-stream backbone that leverages convolutional detail and attention-based context in tandem.

2.3 Attention Mechanisms in Vision Tasks

Attention mechanisms have become ubiquitous in modern vision models (Guo et al., 2022a). In CNN-based networks, specialized attention modules reweight features to emphasize salient information. For example, Squeeze-and-Excitation blocks (Hu et al., 2018) apply channel-wise attention, and convolutional block attention module (Woo et al., 2018) extends this with spatial attention. In parallel, self-attention was first integrated into CNNs by Wang et al. (2018) via the Non-Local Network, which models long-range dependencies across the image. This work showed that capturing global interactions in convolutional features significantly boosts tasks like detection and segmentation. Building on this, purely attention-driven Vision Transformers (ViT) (Dosovitskiy et al., 2020) have been developed for classification and other vision tasks. ViTs tokenize the image and use multi-head self-attention to aggregate information globally, and they have demonstrated "huge potential" by achieving competitive or superior results to CNNs on many benchmarks. Attention is also being used to integrate information across tasks or modalities (Soydaner, 2022; Lu et al., 2023). For example, Lopes et al. (2023) introduce a cross-task attention mechanism in a multi-task framework: they use correlation-guided attention to exchange features pairwise between task-specific branches, which enhances the shared representations for all task. This shows that attention can explicitly model task correlations. Inspired by such ideas, our model employs cross-task multi-head attention in the decoder heads to allow style, artist, and genre predictions to attend to each other's features. In summary, attention modules, whether channel/spatial within a CNN, self-attention in a Transformer, or cross-attention across tasks, are a common tool in vision, enabling more powerful and context-aware feature learning.

3. CTHArt Network

In this Section, we first proposed the CNN-Transformer hybrid backbone, which captures local artistic details and global stylistic semantics and is shared by multi-tasks. Then, a cross-task attention mechanism is introduced between task-specific decoder heads, which explicitly captures inter-task dependencies through learnable feature interactions.

3.1 CNN-Transformer Hybrid Backbone

In this section, we present the architecture of the proposed Convolution-Transformer Hybrid Attention model (CTHArt) for joint artist, style, and genre classification in fine art paintings. The primary goal of our design is to unify local texture learning and global semantic modeling within a single, end-to-end trainable backbone. To achieve this, we construct a dual-branch network composed of a CNN stream and a ViT stream, as shown in Fig. 2. The CNN branch provides strong locality and texture inductive biases, reducingthe burden on the Transformer to learn low-level visual structures from limited data. The Transformer branch thus focuses on modeling long-range semantic dependencies, improving representational efficiency under data-constrained settings. The CNN branch is based on the ResNet-50 architecture, while the Transformer branch adopts the Swin Transformer framework. Both branches are carefully synchronized in terms of feature map resolution and channel dimensions, enabling effective fusion and interaction across stages. Furthermore, the entire backbone is shared among the multiple classification tasks, which facilitates joint feature representation and captures the intrinsic relationships between artistic attributes.

3.1.1 Overall Architecture

The proposed CTHArt backbone consists of two parallel and interactive branches: a CNN stream and a Transformer stream. Each stream processes the input image through a four-stage hierarchical structure, and their respective outputs at each stage are fused bidirectionally to enable the integration of local and global features. Formally, given an input image $I \in \mathbb{R}^{H\times W\times 3}$, the network first downsamples it to a common resolution of $\frac{H}{4} \times \frac{W}{4}$ with 48 channels using convolution and patch embedding operations in the CNN and Transformer branches, respectively. These initial operations ensure that both branches start with feature maps of matching spatial and channel dimensions, which is critical for effective cross-branch interaction. Let $F_1^c, F_2^c, F_3^c, F_4^c$ denote the feature maps from the CNN branch at stages 1 through 4, and $F_1^t, F_2^t, F_3^t, F_4^t$ those from the Transformer branch. At each semantic stage, we perform bidirectional, stage-aligned feature fusion be-

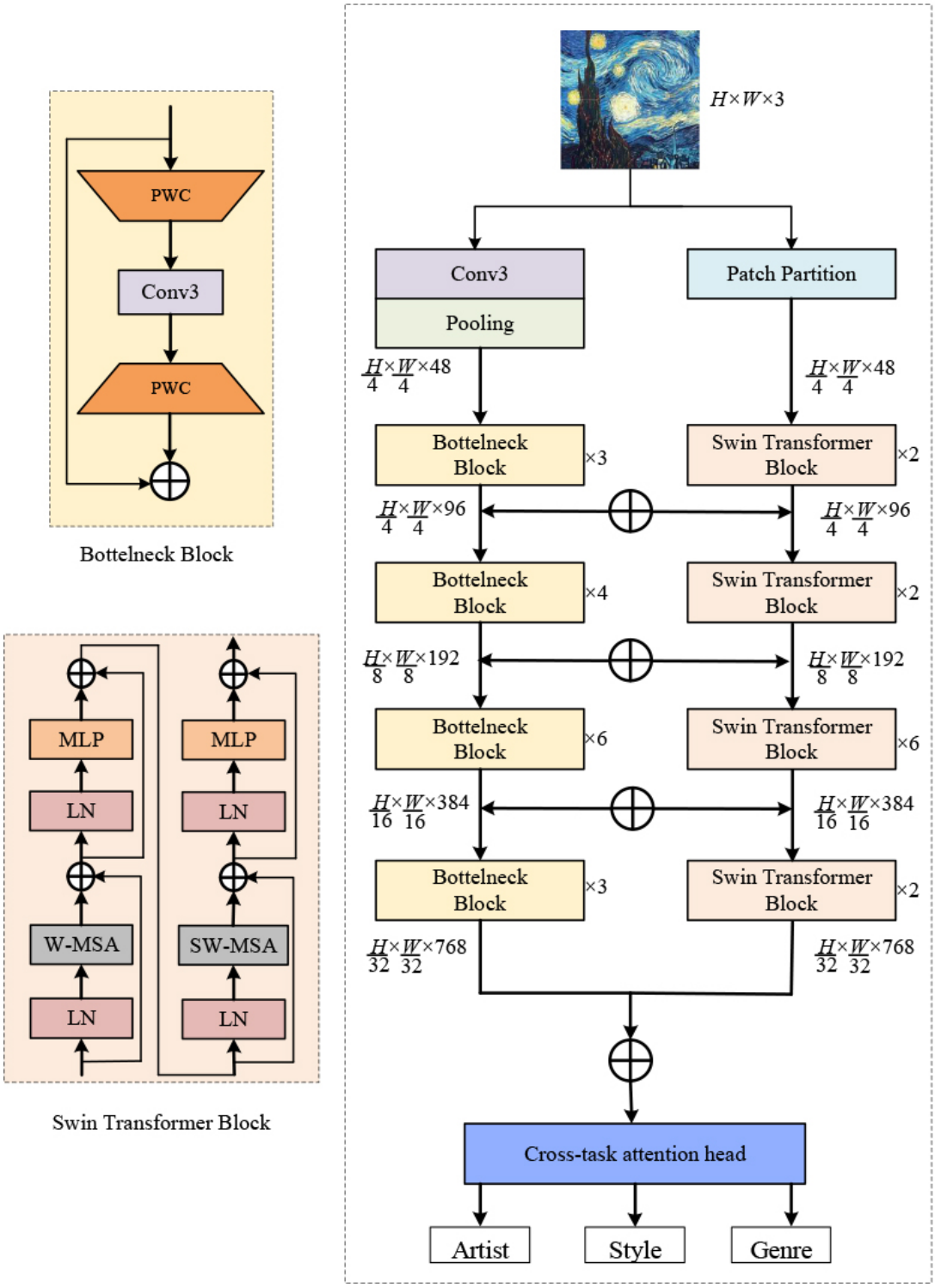

Fig. 2. The architecture of convolutional neural network (CNN)-Transformer hybrid backbone. PWC, Pointwise Convolution; MLP, Multilayer Perceptron; LN, Layer Normalization; W-MSA, Window based Multihead Self-attention; SW-MSA, Shifted windows based Multihead Self-attention.

tween the CNN and Transformer streams. Instead of a single late fusion, local and global representations are recursively integrated prior to further abstraction, enabling progressive refinement of both texture-sensitive and context-aware features. This cross-fusion mechanism facilitates information exchange across the two modalities. Finally, the outputs from both branches at the last stage (F_4^c and F_4^t) are summed and fed into a cross-task attention head, which enables joint prediction across the artist, style, and genre classification tasks.

3.1.2 CNN Branch: ResNet-50 Backbone

The CNN branch follows the structure of ResNet-50 with certain adaptations to match the hybrid design. The input image first passes through a 3×3 convolutional layer followed by a 2×2 max-pooling operation, reducing the spatial resolution from $H \times W$ to $\frac{H}{4} \times \frac{W}{4}$, while projecting the input into a 48-channel feature map. The backbone consists of four sequential stages, each comprising a series of bottleneck blocks. The number of blocks per stage is set to 3, 4, 6, and 3, respectively, mirroring the ResNet-50 design. As shown in Fig. 2, each bottleneck block is composed of three layers: a 1×1 convolution for dimensionality reduction, a 3×3 convolution for spatial feature extraction, and a 1×1 convolution for dimensionality restoration. These blocks facilitate deep feature extraction while maintaining computational efficiency through residual connections. The first block of each stage (except the first stage) performs spatial downsampling via a stride-2 convolution, reducing the feature map resolution by half and simultaneously doubling the number of output channels. This results in feature maps with resolutions and channel sizes as follows: $\frac{H}{4} \times \frac{W}{4} \times 96$, $\frac{H}{8} \times \frac{W}{8} \times 192$, $\frac{H}{16} \times \frac{W}{16} \times 384$, and $\frac{H}{32} \times \frac{W}{32} \times 768$ for stages 1 through 4, respectively.

3.1.3 Transformer Branch: Swin Transformer Backbone

In parallel, the Transformer branch adopts the Swin Transformer architecture, a hierarchical ViT variant designed for scalable image recognition. The Swin Transformer begins by partitioning the input image into non-overlapping patches, each of size 4×4, and embedding them into a 48-dimensional feature space. This results in an initial feature map of size $\frac{H}{4} \times \frac{W}{4} \times 48$, matching the CNN branch. The Swin Transformer consists of four stages, with each stage comprising 2, 2, 6, and 2 Swin Transformer blocks, respectively. Swin Transformer block allows for cross-window connections and enhances global modeling capacity while maintaining computational efficiency. At the first of each stage, the resolution is reduced by half, and the number of channels is doubled, similar to the CNN branch. This hierarchical structure ensures that feature maps at corresponding stages in the CNN and Transformer branches have the same spatial and channel dimensions, facilitating seamless fusion.

3.1.4 Cross-Branch Feature Fusion

To promote mutual learning between local and global representations, we introduce a bidirectional feature fusion mechanism. Specifically, at each stage $s \in 1, 2, 3, 4$ the input to the CNN stage s is the sum of the CNN output from stage s – 1 and the Transformer output from stage s – 1. Similarly, the input to the Transformer stage s is the sum of its own previous output and the CNN output from stage s – 1:

$$\begin{aligned} \tilde{F}_s^c &= F_{s-1}^c + F_{s-1}^t \\ \tilde{F}_s^t &= F_{s-1}^t + F_{s-1}^c \end{aligned} \tag{1}$$

These fused features are then passed through the respective stage's processing blocks. This cross-fusion mechanism introduces a lightweight yet effective inductive bias, enforcing semantic alignment between convolutional and attention-based features. The use of element-wise addition avoids excessive parameterization while ensuring that both branches co-evolve across stages. This strategy enables rich cross-modal interactions and allows the network to simultaneously capture fine-grained local textures (e.g., brushstrokes) and holistic composition cues (e.g., spatial layout, style).

3.1.5 Shared Backbone

A central feature of CTHArt is the use of a shared CNN-Transformer hybrid backbone for all classification tasks. This shared representation is advantageous for multi-task learning, as it allows the model to exploit commonalities and correlations among the tasks of artist attribution, style recognition, and genre identification. For instance, stylistic features often serve as strong priors for inferring genre or even identifying the artist. The outputs from the final stage of both branches (F_4^c and F_4^t) are added and passed to a unified cross-task attention head. This module consists of several task-specific decoder heads—one for each classification task—that interact through a multi-head attention mechanism. Each head attends to the others' features through key-query-value operations, enabling knowledge transfer and promoting consistency across tasks.

The proposed CTHArt network integrates the strengths of CNNs and Transformers in a unified, multitask architecture tailored for art painting categorization. The CNN branch captures local textures and edge details critical for artist and style identification, while the Transformer branch models global relationships and semantic layouts essential for genre recognition. The bidirectional feature fusion and cross-task attention further enhance the model's ability to jointly learn from multiple, interdependent tasks.

3.2 Cross-Task Attention

To capture interdependencies among tasks, we introduce a Cross-Task Attention Head (CTAHead), which operates on the global representation obtained from the shared backbone as shown in Fig. 3 (Taking MultitaskPainting100k dataset as an example). The final feature map from the backbone is globally averaged, yielding a compact representation $F \in \mathbb{R}^{1\times1\times768}$. This feature is projected via a fully connected layer into a 1674-dimensional vector $O \in \mathbb{R}^{1\times1\times1674}$, where O is partitioned into three sub-vectors corresponding to the three tasks:

$$O = [O^a, O^s, O^g] \tag{2}$$

with $O^a \in \mathbb{R}^{1\times 1508}$ (artist), $O^s \in \mathbb{R}^{1\times 125}$ (style), and $O^g \in \mathbb{R}^{1\times 41}$ (genre).

To model task dependencies, we employ a cross-task attention mechanism based on parallel task heads with pairwise attention interactions. The three tasks (artist, style, and genre) are predicted simultaneously, and each task head attends to the feature representations of the other two tasks to capture inter-task correlations. Each task attends to the other two through pairwise cross-attention, forming a parallel and mutually interactive multitask prediction structure. The final output of each head is passed through a fully connected layer followed by a softmax operation to yield the artist classification probabilities. The calculation process of artist decoder head is as follows:

Each sub-vector is projected to task-specific query, key, and value vectors:

$$\begin{aligned} Q^a &= O^a W_a^Q, K^a = O^a W_a^K, V^a = O^a W_a^V \\ Q^s &= O^s W_s^Q, K^s = O^s W_s^K, V^s = O^s W_s^V \\ Q^g &= O^g W_g^Q, K^g = O^g W_g^K, V^g = O^g W_g^V \end{aligned} \tag{3}$$

We first perform attention from the artist task to the style task:

$$O_1^a = soft\max\left(\frac{Q^a (K^s)^T}{\sqrt{d}}\right) V^s \tag{4}$$

The intermediate output Q1 a is then projected:

$$Q_1^a = O_1^a W_{a1}^Q \tag{5}$$

Next, artist queries attend to the genre features:

$$O_2^a = soft\max\left(\frac{Q_1^a (K^g)^T}{\sqrt{d}}\right) V^g \tag{6}$$

The intermediate output Q1 a is then projected:

$$Q_2^a = O_2^a W_{a2}^Q \tag{7}$$

The final artist query attends to its own keys and values:

$$O_3^a = soft\max\left(\frac{Q_2^a (K^a)^T}{\sqrt{d}}\right) V^a \tag{8}$$

This final artist representation is passed through a fully connected layer and softmax activation.

The similar computation flow is applied for style and genre predictions, with task roles rotated accordingly. This structured cross-attention mechanism enables semantic communication across heads, reinforcing predictions with inter-task context. For instance, an expressionist style feature may support the prediction of a 20th-century genre or a specific artist known for that movement.

This hierarchical attention structure is mirrored for the style and genre heads, with each task attending to the others in a cyclic order. Through this cross-task attention, the decoder heads dynamically exchange semantic information, capturing interdependencies among artistic attributes. This design enhances prediction performance by allowing the network to exploit subtle correlations across tasks, such as stylistic indicators that hint at an artist's identity or genre-related elements that inform stylistic decisions.

4. Experiments and Results

4.1 Dataset and Training Detail

To comprehensively evaluate the performance of the proposed CTHArt model, we con-duct experiments on three widely used benchmark datasets in the domain of computational art analysis: Painting-91 (Khan et al., 2014), WikiArt (Tan et al., 2018), and MultitaskPainting100k (Bianco et al., 2019). These datasets are selected for their diversity in artistic content, their inclusion of multiple classification tasks (e.g., artist, style, genre), and their adoption in previous literature, which enables fair comparisons with existing approaches.

4.1.1 Painting-91

The Painting-91 dataset is one of the earliest curated datasets used in computer vision-based art analysis. It consists of 4266 images from 91 different artists. The artworks cover a variety of painting styles and time periods, making the dataset suitable primarily for artist classification tasks. The number of samples per artist is not uniformly distributed, leading to a moderate class imbalance. This dataset has been extensively used in artist identification tasks and remains a standard benchmark for evaluating model performance in artist attribution.

4.1.2 WikiArt

The WikiArt dataset is a large-scale collection of fine art images compiled from the WikiArt.org repository. It includes over 80,000 images from thousands of artists, spanning various styles, genres, and centuries. For the purpose

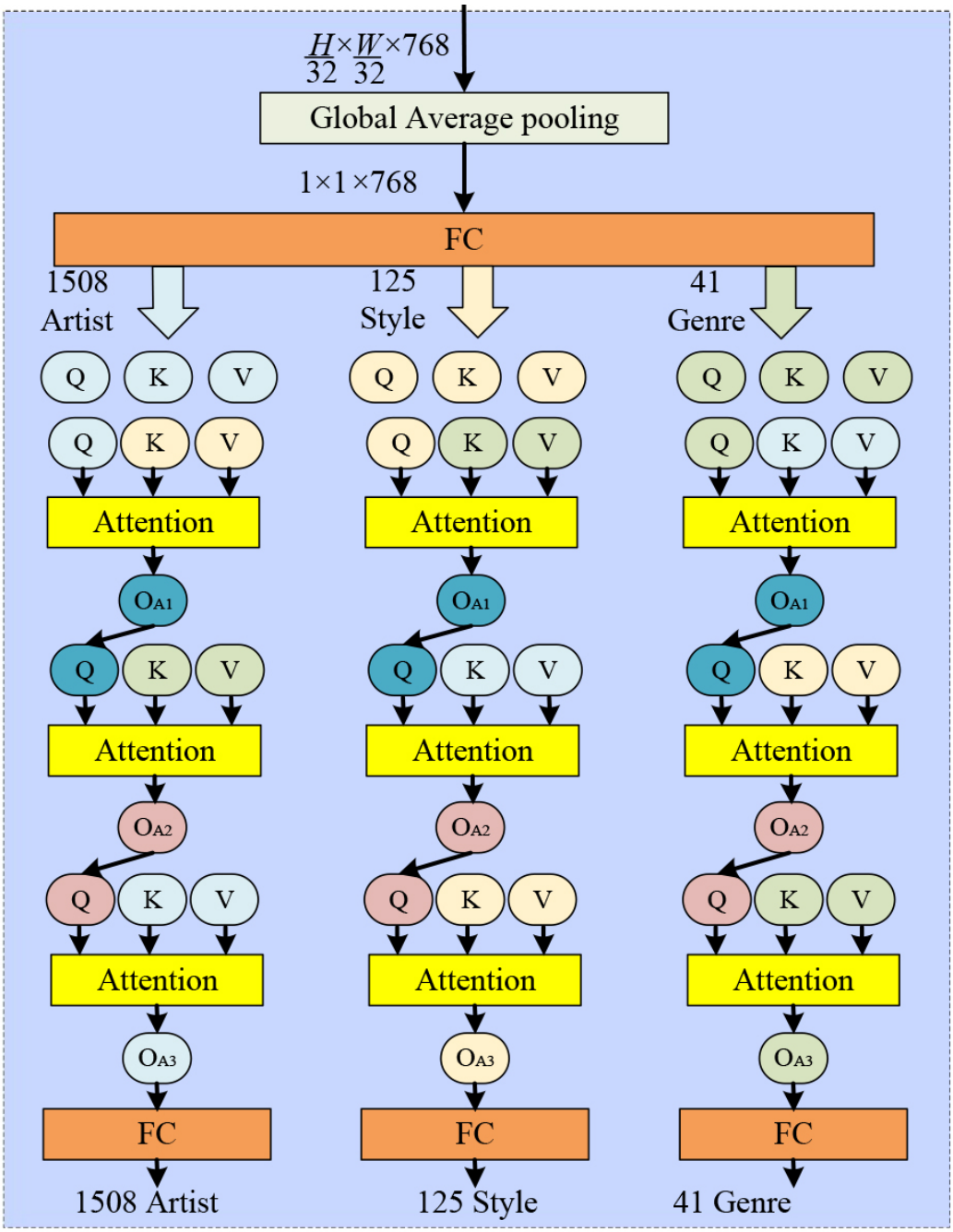

Fig. 3. Architecture of the Cross-Task Attention Head. Three task-specific heads (artist, style, genre) operate in parallel and perform pairwise cross-attention with each other.

of consistent benchmarking, we follow the preprocessed version released by previous works, which consists of a curated subset with three label types: artist, style, and genre. Specifically, we adopt a version containing approximately 65,000 paintings with complete labels across all three tasks. The style labels include 27 classes, such as Impressionism, Expressionism, and Baroque, while the genre labels include 10 categories like Portrait, Landscape, and Abstract. Artist labels in this subset span 1195 artists. WikiArt dataset supports multitask classification and allows joint learning of multiple visual attributes. However, due to the large number of artist classes and data imbalance across artists and styles, training robust models on this dataset remains challenging.

4.1.3 MultitaskPainting100k

The MultitaskPainting100k dataset is a recent large-scale benchmark introduced to facilitate multi-label and multitask learning in computational art understanding. It contains 99,816 images labeled with three hierarchical attributes: 1508 artists, 125 styles, and 41 genres. This dataset is significantly larger and more comprehensive than

both Painting-91 and WikiArt, and it introduces several challenges, including extreme label sparsity, high inter-class correlation, and severe class imbalance. Each image is annotated with all three types of labels, making this dataset particularly suitable for evaluating the effectiveness of joint representation learning and cross-task interactions. The large label space (especially for artist classification) poses a challenge for both representation capacity and regularization.

The three datasets used in our study cover complementary aspects of fine art classification. Painting-91 emphasizes artist identification under limited data; WikiArt offers moderate-scale multitask classification with diverse artistic content; and MultitaskPainting100k challenges models with extreme-scale multitask learning. This comprehensive setup allows us to evaluate both the accuracy and scalability of the proposed CTHArt architecture.

4.1.4 Training Detail

To train the proposed CTHArt model effectively across multiple tasks, we adopt a consistent and robust training strategy based on best practices in deep visual representation learning. All images are resized to 224×224 pixels before training. Data augmentation plays a crucial role in improving generalization and robustness to artistic variability. During training, we employ standard augmentation techniques, including random horizontal flipping, random cropping, color jittering, and normalization using ImageNet statistics. The model is trained using the Adam optimizer, which provides stable convergence in multitask learning scenarios. The initial learning rate is set to 0.1, and we apply cosine annealing to gradually reduce the learning rate over the course of training, reaching a final value of 0.0001. This schedule helps the model escape sharp local minima early in training and settle into flatter minima, which typically generalize better. We train the model with a batch size of 128 on NVIDIA Tesla V100 GPUs (Santa Clara, CA, USA). The multitask loss is calculated as the weighted sum of the individual cross-entropy losses for the artist, style, and genre classifiers. Weights are selected empirically to balance gradients and ensure that no single task dominates the optimization process. This training setup ensures that the CTHArt model is exposed to a diverse range of visual patterns while learning to generalize across three interdependent classification tasks.

4.2 Results and Analysis

We compare our approach against a broad range of state-of-the-art CNN models [the accuracies of ResNet (He et al., 2016), Res2Net (Gao et al., 2021), ResNeXt (Xie et al., 2017), RegNet (Xu et al., 2023), ResNeSt (Zhang et al., 2022), Efficient-Net (Tan and Le, 2019) are reported in (Zhao et al., 2021)], Transformer-based models [ViT (Dosovitskiy et al., 2020), Swin-Transformer (Liu et al., 2021)], hybrid networks [CMT (Guo et al., 2022b), CoAtNet (Dai et al., 2021)], and report performance across artist, style, and genre classification tasks.

4.2.1 Results on Painting-91

Painting-91 is a relatively small-scale dataset, primarily used to evaluate artist and style classification. Table 1 summarizes the performance of various methods. Our model achieves the highest accuracy on both artist and style classification tasks, with 73.25% and 80.84% respectively. Compared with traditional CNN backbones such as ResNet (58.92%, 66.54%) and Res2Net (58.81%, 70.13%), CTHArt shows significant improvements, highlighting the benefits of integrating Transformer-based global context modeling. Even against more advanced architectures like RegNetX and EfficientNet, our model demonstrates superior results. EfficientNet, known for its scaling efficiency, achieves 71.27% artist accuracy and 79.23% style accuracy, which are both surpassed by our approach. The improvement in artist classification, which typically depends on fine-grained local texture and brushstroke details, suggests that our CNN branch effectively captures discriminative patterns. Meanwhile, the performance gain in style classification implies that the Transformer stream and attention modules successfully encode higher-level semantics and global composition cues.

4.2.2 Results on WikiArt and MultitaskPainting100k

Table 2 presents the results of our model and various baselines on the WikiArt and MultitaskPainting100k datasets. These datasets are larger and more complex than Painting-91, incorporating all three classification tasks: artist, style, and genre.

The WikiArt dataset provides a large-scale and diverse benchmark for multi-facet art classification, covering artist, style, and genre simultaneously. As shown in Table 2, our CTHArt model achieves the best overall performance among all compared methods, reaching a mean accuracy of 81.18%, outperforming both pure CNN, pure Transformer, and recent CNN-Transformer hybrid architectures. Compared with strong CNN baselines such as ResNeSt (78.69%) and EfficientNet (79.65%), our model improves the mean accuracy by 2.49% and 1.53%, respectively. This improvement indicates that incorporating global self-attention modeling alongside convolutional inductive bias yields more discriminative representations for complex artistic attributes. In particular, CNN backbones tend to perform competitively on artist classification due to their strength in capturing local texture patterns, but they are relatively weaker in style recognition, which requires more global compositional and semantic understanding.

Transformer-based models such as ViT and Swin-Transformer achieve strong performance, but at higher computational cost. ViT requires 55.4 Giga FLOPs, nearly eight times that of our model, while still yielding lower mean accuracy (78.99%). Swin-Transformer improves efficiency but remains inferior to our method in all three tasks.

Table 1. Results on Painting-91.

Method	(Chu and Wu, 2018)	EfficientNet	ResNet	Res2Net	RegNetX	ResNetSt	Ours
Artist	64.32	71.27	58.92	58.81	65.44	60.07	73.25
Style	78.27	79.23	66.54	70.13	73.35	62.96	80.84
Mean	71.30	75.25	62.73	64.47	69.40	61.52	77.05

Table 2. Results on WikiArt and MultitaskPainting100k.

Dataset	Method	FLOPs	Artist	Style	Genre	Mean
WikiArt	(Zhong et al., 2020)	-	88.38	58.99	76.27	74.55
	(Cetinic et al., 2018)	-	81.94	56.43	77.60	71.99
	ResNet	4.2 G	88.03	65.76	76.69	76.83
	Res2Net	4.2 G	88.14	65.97	76.54	76.88
	ResNeXt	4.2 G	88.29	66.62	76.74	77.22
	RegNetX	3.2 G	89.08	67.10	76.63	77.60
	RegNetY	4.0 G	76.29	69.51	77.41	74.40
	ResNetSt	4.3 G	88.17	69.97	77.94	78.69
	EfficientNet	4.2 G	91.73	69.19	78.03	79.65
	ViT	55.4 G	90.65	68.49	77.84	78.99
	Swin-Transformer	8.7 G	91.86	69.02	78.13	79.67
	CMT-B	9.3 G	92.71	69.46	78.79	80.32
	CoAtNet	8.4 G	93.03	68.97	79.21	80.40
	Ours	7.3 G	93.24	70.65	79.64	81.18
MultitaskPainting100k	(Bianco et al., 2019)	-	56.50	57.20	63.60	59.10
	ResNet	4.2 G	59.05	59.41	65.77	61.41
	Res2Net	4.2 G	61.28	60.81	66.31	62.80
	ResNeXt	4.2 G	59.93	60.73	66.27	62.31
	RegNetX	3.2 G	59.97	60.60	65.99	62.19
	RegNetY	4.0 G	61.22	62.80	66.82	63.61
	ResNetSt	4.3 G	62.78	62.64	67.83	64.42
	EfficientNet	4.2 G	65.50	63.15	66.99	65.21
	ViT	17.6 G	66.05	64.32	68.02	66.13
	Swin-Transformer	8.7 G	66.32	64.29	68.14	66.25
	CMT-B	9.3 G	67.32	64.18	67.93	66.48
	CoAtNet	8.4 G	66.89	65.13	68.32	66.78
	Ours	7.3 G	67.49	65.07	68.99	67.18

This demonstrates that a carefully designed hybrid backbone with staged cross-branch fusion can achieve better accuracy-efficiency trade-offs than standalone Transformer architectures.

When compared with recent hybrid models such as CMT-B and CoAtNet, our approach still shows consistent gains. CTHArt achieves the highest accuracy in all three tasks, surpassing CoAtNet by 0.78% in mean accuracy while using lower computational cost (7.3 G vs. 8.4 Giga FLOPs). The most notable gain appears in style classification, where global stylistic semantics and cross-task cues are especially important. This suggests that the proposed cross-task attention head effectively transfers complementary information between artist, style, and genre branches, leading to more semantically consistent predictions.

MultitaskPainting100k is a more challenging large-scale multitask benchmark with a very large artist label space and strong class imbalance. As reported in Table 2, 400 the proposed CTHArt model again achieves the best overall performance, with a mean accuracy of 67.18%, outperforming all compared CNN, Transformer, and hybrid baselines.

Compared with the original multitask CNN model (Bianco et al., 2019), our method improves the mean accuracy by more than 8 percentage points, confirming the effectiveness of hybrid representation learning and explicit cross-task interaction under extreme label cardinality. Traditional CNN backbones such as ResNet, Res2Net, and RegNet variants achieve mean accuracies in the range of 61–64%, indicating limited capacity in modeling long-range structure and cross-facet semantics when tasks are learned jointly.

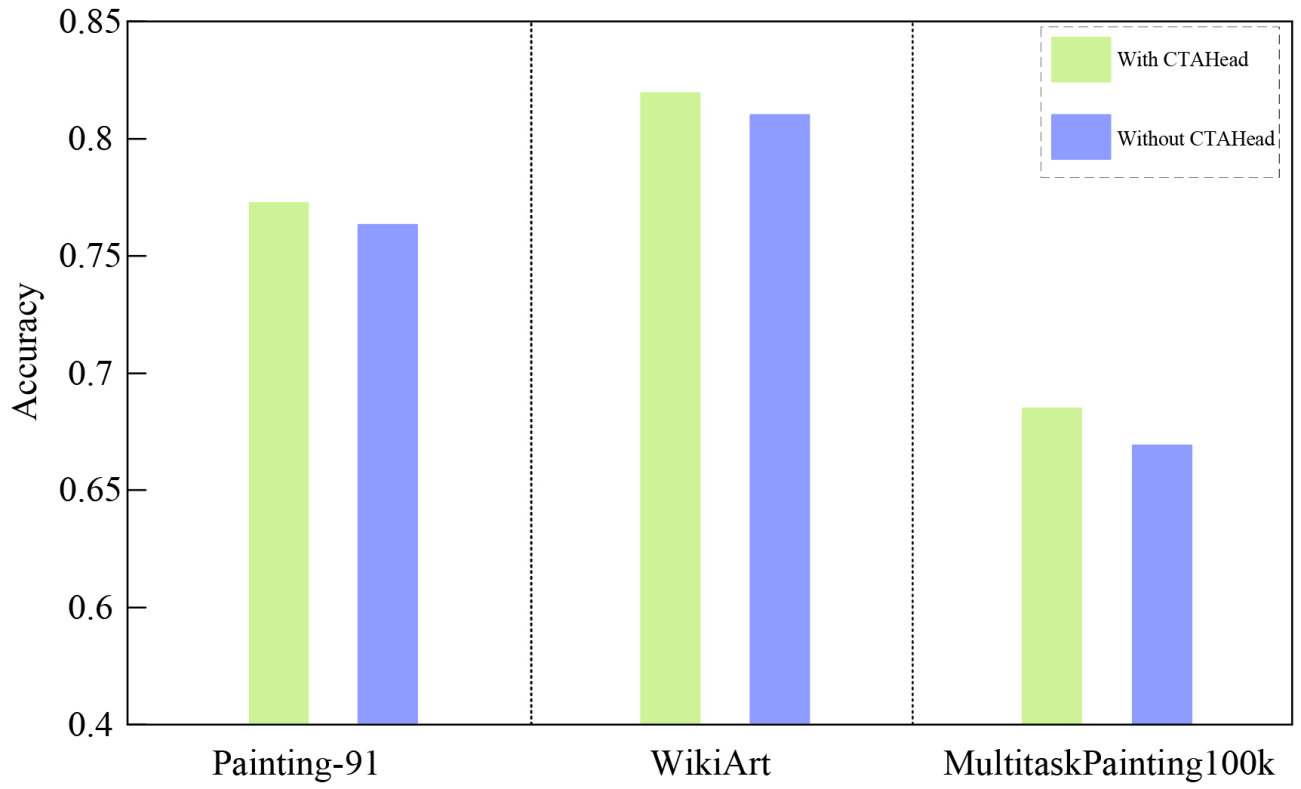

Fig. 4. The results of ablation study. CTAHead, Cross-Task Attention Head.

Transformer-based models provide stronger global modeling and show clear gains over CNNs. ViT and Swin-Transformer reach mean accuracies of 66.13% and 66.25%, respectively. However, our model still surpasses them while requiring fewer FLOPs than ViT and slightly less than Swin. This suggests that combining convolutional texture priors with hierarchical Transformer features is more data-efficient than relying on attention mechanisms alone, especially in fine-grained art datasets where training samples per class can be limited.

Among hybrid architectures, CMT-B and CoAtNet also show competitive performance, with mean accuracies of 66.48% and 66.78%. Our model further improves the mean score to 67.18%, achieving the best artist accuracy (67.49%) and genre accuracy (68.99%). Overall, the WikiArt and MultitaskPainting100k results verify that the proposed multitask hybrid architecture not only improves accuracy across all facets but also maintains favorable computational efficiency. Notably, the performance gains on MultitaskPainting100k are consistent but more moderate than those on WikiArt. This behavior is expected due to the heavier label imbalance and larger class space, which increases task difficulty and reduces the margin between strong models. Nevertheless, the proposed method still delivers the top overall performance with controlled computational complexity.

In addition to accuracy, we further evaluate our model using macro F1-score on datasets with strong class imbalance. On WikiArt and MultitaskPainting100k, our method achieves F1-scores of 0.74 and 0.62, respectively. Since prior published methods on these benchmarks report only accuracy and do not provide precision/recall or F1 statistics, direct F1 comparison is not available. We therefore report F1 results for our model to provide complementary and more imbalance-aware evaluation while keeping accuracy-based comparisons for fairness with existing literature.

4.3 Ablation Study

To better understand the contribution of the proposed CTAHead, we conduct an ablation study by comparing the full CTHArt model with a reduced variant where the CTAHead is removed. In the ablated model, the global feature vector obtained from the backbone is directly passed to three independent fully connected layers for artist, style, and genre classification, without any task interaction or attention-based reasoning. All other components (hybrid backbone, training protocols) remain identical. As shown in Fig. 4, the mean accuracy was improved by more than 1% on three datasets, with the help of CTAHead.

Ablation experiments demonstrate that removing CTAHead leads to notable performance drops across all tasks, confirming its critical role in the architecture. By embedding structured attention across artist, style, and genre representations, CTAHead elevates the model's capacity for nuanced interpretation in the artistic domain.

5. Conclusions

In this work, we presented CTHArt, a CNN-Transformer hybrid multitask framework for fine art painting categorization across artist, style, and genre facets. The proposed architecture integrates a dual-branch hybrid backbone with stage-wise cross-branch fusion, enabling the model to jointly capture fine-grained local artistic details and global semantic structure. On top of the shared backbone, we introduced a Cross-Task Attention Head to improve accuracy by inter-task dependencies.

Extensive experiments on Painting-91, WikiArt, and MultitaskPainting100k demonstrate that the proposed model consistently outperforms strong CNN, Transformer, and recent hybrid baselines, achieving state-of-the-art accuracy with competitive computational cost. The gains are especially evident in multitask settings with large label spaces and class imbalance, confirming the benefit of hybrid representation learning and cross-task semantic interaction. Beyond accuracy gains, the proposed framework also provides a practical technical pathway for AI-assisted knowledge organization in large-scale art repositories, supporting coordinated facet assignment and semantically consistent metadata construction. Future work will explore lighter-weight hybrid designs, more advanced task-relation modeling strategies, and extensions to additional art attributes and multimodal cultural heritage data.

Availability of Data and Materials

All data reported in this paper will be shared by the corresponding author upon resonable request.

Author Contributions

LW: Formal Analysis, Methodology, Software, Writing - Review & Editing, Validation. QL: Funding Acquisition, Supervision, Writing - Original Draft, Visualization, Resources. DW: Project Administration, Investigation, Conceptualization, Data Curation. DW reviewed the paper critically for important intellectual content. All authors contributed to editorial changes in the manuscript. All authors have reviewed the final version of the manuscript and have agreed to its publication; all authors take responsibility for the study.

Acknowledgment

Not applicable.

Funding

This research was funded by Shandong Province Cultural and Tourism Research Project (Project No. 24WL(Y)91).

Conflicts of Interest

The authors declare no conflicts of interest.

References

Arshad T, Zhang J, Ullah I. A hybrid convolution transformer for hyperspectral image classification. European Journal of Remote Sensing. 2024; 57: 2330979. https://doi.org/10.1080/22797254.2024.2330979.

Bagchi M. A Large Scale, Knowledge Intensive Domain Development Methodology. Knowledge Organization. 2021a; 48: 8–23. https://doi.org/10.5771/0943-7444-2021-1-8.

Bagchi M. Towards knowledge organization ecosystem (KOE). Cataloging & Classification Quarterly. 2021b; 59: 740–756. https://doi.org/10.1080/01639374.2021.1998282.

Bianco S, Mazzini D, Napoletano P, Schettini R. Multitask painting categorization by deep multibranch neural network. Expert Systems with Applications. 2019; 135: 90–101. https://doi.org/10.1016/j.eswa.2019.05.036.

Cetinic E, Lipic T, Grgic S. Fine-tuning Convolutional Neural Networks for fine art classification. Expert Systems with Applications. 2018; 114: 107–118. https://doi.org/10.1016/j.eswa.2018.07.026.

Chen D, Miao D, Zhao X. Hyneter: Hybrid Network Transformer for Multiple Computer Vision Tasks. IEEE Transactions on Industrial Informatics. 2024; 20: 8773–8785. https://doi.org/10.1109/TII.2024.3367043.

Chu WT, Wu YL. Image style classification based on learnt deep correlation features. IEEE Transactions on Multimedia. 2018; 20: 2491–2502. https://doi.org/10.1109/TMM.2018.2801718.

Dai Z, Liu H, Le QV, Tan M. Coatnet: Marrying convolution and attention for all data sizes. Advances in Neural Information Processing Systems. 2021; 34: 3965–3977. https://doi.org/10.48550/arXiv.2106.04803.

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929. 2020. https://doi.org/10.48550/arXiv.2010.11929. (preprint)

Fang J, Lin H, Chen X, Zeng K. A hybrid network of cnn and transformer for lightweight image super-resolution. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1103–1112). 2022. https://doi.org/10.1109/CVPRW56347.2022.00119.

Gao S, Cheng M, Zhao K, Zhang X, Yang M, Torr P. Res2Net: a New Multi-Scale Backbone Architecture. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021; 43: 652–662. https://doi.org/10.1109/TPAMI.2019.2938758.

Gao Z, Chen J, Liu Y, Jin Y, Tian D. A systematic survey on human pose estimation: upstream and downstream tasks, approaches, lightweight models, and prospects. Artificial Intelligence Review. 2025; 58: 68. https://doi.org/10.1007/s10462-024-11060-2.

Giunchiglia F, Bagchi M. From knowledge representation to knowledge organization and back. Wisdom, Well-Being, Win-Win. In International Conference on Information (pp. 270–287). Springer Nature Switzerland: Cham. 2024. https://doi.org/10.1007/978-3-031-57850-2_20.

Giunchiglia F, Dutta B, Maltese AV. From Knowledge Organization to Knowledge Representation. Knowledge Organization. 2014; 41: 44–56. https://doi.org/10.5771/0943-7444-2014-1-44.

González-Martín C, Carrasco M, Wachter Wielandt TG. Detection of Emotions in Artworks Using a Convolutional Neural Network Trained on Non-Artistic Images: a Methodology to Reduce the Cross-Depiction Problem. Empirical Studies of the Arts. 2024; 42: 38–64. https://doi.org/10.1177/02762374231163481.

Guo J, Han K, Wu H, Tang Y, Chen X, Wang Y, et al. Cmt: Convolutional neural networks meet vision transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 12175–12185). 2022a. https://doi.org/10.48550/arXiv.2107.06263.

Guo M, Xu T, Liu J, Liu Z, Jiang P, Mu T, et al. Attention mechanisms in computer vision: a survey. Computational Visual Media. 2022b; 8: 331–368. https://doi.org/10.1007/s41095-022-0271-y.

He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770–778). 2016. https://doi.org/10.48550/arXiv.1512.03385.

Hjørland B. What is Knowledge Organization (KO)? Knowledge Organization. 2008; 35: 86–101. https://doi.org/10.5771/0943-7444-2008-2-3-86.

Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132–7141). 2018. https://doi.org/10.48550/arXiv.1709.01507.

Khan FS, Beigpour S, van de Weijer J, Felsberg M. Painting-91: a large scale database for computational painting categorization. Machine Vision and Applications. 2014; 25: 1385–1397. https://doi.org/10.1007/s00138-014-0621-6.

Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012–10022). 2021. https://doi.org/10.48550/arXiv.2103.14030.

Liu X. Deep Art Understanding: A Unified Framework for Multi-Task Art Analysis with Cross-Modal Attention. In Proceedings of the 2024 International Conference on Artificial Intelligence, Digital Media Technology and Interaction Design (pp. 566–571). 2024. https://doi.org/10.1145/3726010.3726099.

Long H. Hybrid design of CNN and vision transformer: A review. In Proceedings of the 2024 7th International Conference on Computer Information Science and Artificial Intelligence (pp. 121–127). 2024. https://doi.org/10.1145/3703187.3703208.

Lopes I, Vu T, de Charette R. Cross-task attention mechanism for dense multi-task learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 2329–2338). 2023. https://doi.org/10.48550/arXiv.2206.08927.

Lu S, Liu M, Yin L, Yin Z, Liu X, Zheng W. The multimodal fusion in visual question answering: a review of attention mechanisms. PeerJ Computer Science. 2023; 9: e1400. https://doi.org/10.7717/peerj-cs.1400.

Shah SA, Taj I, Usman SM, Hassan Shah SN, Imran AS, Khalid S. A hybrid approach of vision transformers and CNNs for detection of ulcerative colitis. Scientific Reports. 2024; 14: 24771. https://doi.org/10.1038/s41598-024-75901-4.

Soydaner D. Attention mechanism in neural networks: where it comes and where it goes. Neural Computing and Applications. 2022; 34: 13371–13385. https://doi.org/10.1007/s00521-022-07366-3.

Specker E, Douda M, Leder H. How do we Understand Artworks? Exploring the Role of Artwork Inherent Features in Art Processing. Empirical Studies of the Arts. 2024; 42: 469–497. https://doi.org/10.1177/02762374231201074.

Strezoski G, Worring M. Omniart: a large-scale artistic benchmark. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM). 2018; 14: 1–21. https://doi.org/10.1145/3273022.

Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning (pp. 6105–6114). PMLR. 2019. https://doi.org/10.48550/arXiv.1905.11946.

Tan WR, Chan CS, Aguirre HE, Tanaka K. Improved ArtGAN for Conditional Synthesis of Natural Image and Artwork. IEEE Transactions on Image Processing. 2018; 28: 394–409. https://doi.org/10.1109/TIP.2018.2866698.

Tian T, Nan F. A Multitask Convolutional Neural Network for Artwork Appreciation. Mobile Information Systems. 2022; 2022: 8804711. https://doi.org/10.1155/2022/8804711.

Ugail H, Stork DG, Edwards H, Seward SC, Brooke C. Deep transfer learning for visual analysis and attribution of paintings by Raphael. Heritage Science. 2023; 11: 1–5. https://doi.org/10.1186/s40494-023-01094-0.

Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7794–7803). 2018. https://doi.org/10.48550/arXiv.1711.07971.

Woo S, Park J, Lee J, Kweon IS. CBAM: Convolutional Block Attention Module. In Proceedings of the European conference on computer vision (ECCV) (pp. 3–19). 2018. https://doi.org/10.48550/arXiv.1807.06521.

Wu H, Xiao B, Codella N, Liu M, Dai X, Yuan L, et al. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 22–31). 2021. https://doi.org/10.48550/arXiv.2103.15808.

Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1492–1500). 2017. https://doi.org/10.48550/arXiv.1611.05431.

Xu J, Pan Y, Pan X, Hoi S, Yi Z, Xu Z. RegNet: Self-Regulated Network for Image Classification. IEEE Transactions on Neu-

ral Networks and Learning Systems. 2023; 34: 9562–9567. https://doi.org/10.1109/TNNLS.2022.3158966.

Xu R, Hsu Y, Wang X. Examining Artificial Intelligence Anxiety in the Context of Anthropocentrism and the Art Turing Test. Empirical Studies of the Arts. 2025. https://doi.org/10.1177/02762374251353671.

Yang B, Xiang X, Kong W, Peng Y, Yao J. Adaptive multi-task learning using lagrange multiplier for automatic art analysis. Multimedia Tools and Applications. 2022; 81: 3715–3733. https://doi.org/10.1007/s11042-021-11360-7.

Yang J, Intan Raihana Ruhaiyem N, Zhou C. A 3M-Hybrid Model for the Restoration of Unique Giant Murals: a Case Study on the Murals of Yongle Palace. IEEE Access. 2025; 13: 38809–38824. https://doi.org/10.1109/ACCESS.2025.3542320.

Zhang H, Wu C, Zhang Z, Zhu Y, Lin H, Zhang Z, et al. Resnest: Split-attention networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2736–2746). 2022. https://doi.org/10.1109/CVPRW56347.2022.00309

Zhao Q, Zhang R. Classification of painting styles based on the difference component. Expert Systems with Applications. 2025; 259: 125287. https://doi.org/10.1016/j.eswa.2024.125287.

Zhao W, Zhou D, Qiu X, Jiang W. Compare the performance of the models in art classification. Plos one. 2021; 16: e0248414. https://doi.org/10.1371/journal.pone.0248414.

Zhao Z, Bai H, Zhang J, Zhang Y, Xu S, Lin Z, et al. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5906–5916). 2023. https://doi.org/10.48550/arXiv.2211.14461.

Zhong S, Huang X, Xiao Z. Fine-art painting classification via two-channel dual path networks. International Journal of Machine Learning and Cybernetics. 2020; 11: 137–152. https://doi.org/10.1007/s13042-019-00963-0.