Research Square

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

HierWoodNet: HierStyle Multi-scale Semantic Style Transfer Method for Chinese Woodblock Painting

Jie Qiao
Northwest University
Fangqing Fan
308912764@qq.com

Northwest University

Article

Keywords:

Posted Date: February 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8532155/v1

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

Additional Declarations: No competing interests reported.

HierWoodNet: HierStyle Multi-scale Semantic Style Transfer Method for Chinese Woodblock Painting

Jie Qiao[1,2] **and Fangqing Fan**[3,*]

[1]School of Electronic Information, Northwest University, Xi'an, China
[2]Yulin Science and Technology Bureau,Yu'lin, China
[3]Network and Data Center, Northwest University, Xi'an, China
[*]corresponding.308912764@qq.com

ABSTRACT

Woodblock painting, a traditional Chinese intangible cultural heritage, is known for its vivid colors and distinctive textures. However, modern lifestyles and declining cultural inheritance have led to risks of style loss and discontinuity. AI-based image style transfer offers a promising avenue for preserving and creating woodblock art. However, existing methods struggle with semantic feature extraction, style consistency, and color modeling in this domain. To address these challenges, we propose a hierarchical woodblock style transfer network. Our approach introduces a Content-Aware Positional Encoder based on the Transformer architecture, which enhances multi-scale feature representation and global structural perception. This ensures faithful content preservation during style transfer. We further present a HierStyle Encoder that learns rich, hierarchical style features from data, enabling fine-grained control over the transfer process and improving semantic and textural coherence. A progressive decoder integrates multi-stage upsampling with cross-layer connections to preserve semantic structure while injecting detailed style features. Finally, the ChromaNova Color Decoder dynamically optimizes color representation through a multi-dimensional color loss function that considers both color contrast and distribution differences. Extensive qualitative and quantitative experiments demonstrate that our method outperforms current general-purpose style transfer approaches on multi-color woodblock style transfer tasks. Our dataset is available at https://github.com/LPDLG/HierWoodNet.

1 Introduction

Woodblock painting, an integral component of China's intangible cultural heritage, originated during the Tang Dynasty and reached its zenith in the Ming and Qing dynasties. This folk art form integrates wood carving and printing techniques to produce visually distinctive works, exemplified by regional schools such as Yangliuqing and Taohuawu—traditionally employed in festive decorations and cultural communication[1]. The creation process entails multiple meticulous stages: drafting, block preparation, carving, inking, and rubbing, each requiring exceptional artisanal skill[2]. Stylistically, Chinese woodblock prints are characterized by contour-driven modeling, symbolic use of color, balanced composition between positive and negative space, and precise multi-block color registration[3,4].

However, industrialization and digital media have marginalized this tradition. Fewer than twenty national-level inheritors remain, and many regional schools are at risk of disappearing. Preserving and revitalizing this art form is thus urgent. Traditional stylization methods evolved from non-photorealistic rendering (NPR)[5] to CNN-based style transfer[6], with advances in online[7,8] and offline[9–11] generation. Yet, these methods struggle with Woodblock painting style due to three main challenges:

• **Insufficient global semantic feature extraction.** Traditional woodblock painting are limited in number and exhibit diverse styles. Existing datasets often fail to cover the full range of style characteristics, posing a challenge for models to effectively learn style representations during training. Moreover, most current style transfer methods focus on local details such as color and texture, neglecting global semantic information. This leads to poor preservation of structural content in the generated images.

• **Difficulty in maintaining style consistency.** Woodblock painting typically contain unique texture distributions, complex layouts, and abstract semantic expressions, which are highly subjective. As a result, existing methods struggle to maintain semantic coherence in style transfer, making it difficult to achieve smooth structural transitions and consistent texture fusion.

• **Limited color-semantic modeling capability.** During the creation of woodblock painting, visual effects such as edge diffusion and pigment bleeding are common. Most existing style transfer methods rely on pixel-wise reconstruction loss or perceptual loss for optimization, which are not well-suited for capturing the high-contrast, concentrated color distributions, and hue-sensitive characteristics of multi-color woodblock painting.

To address the above challenges, we propose a **Hier**Style Multi-scale Semantic Style Transfer Method for Chinese **Wood**block Painting. As shown in the Fig. 1, our model achieving high-quality and expressive style transfer for multi-color

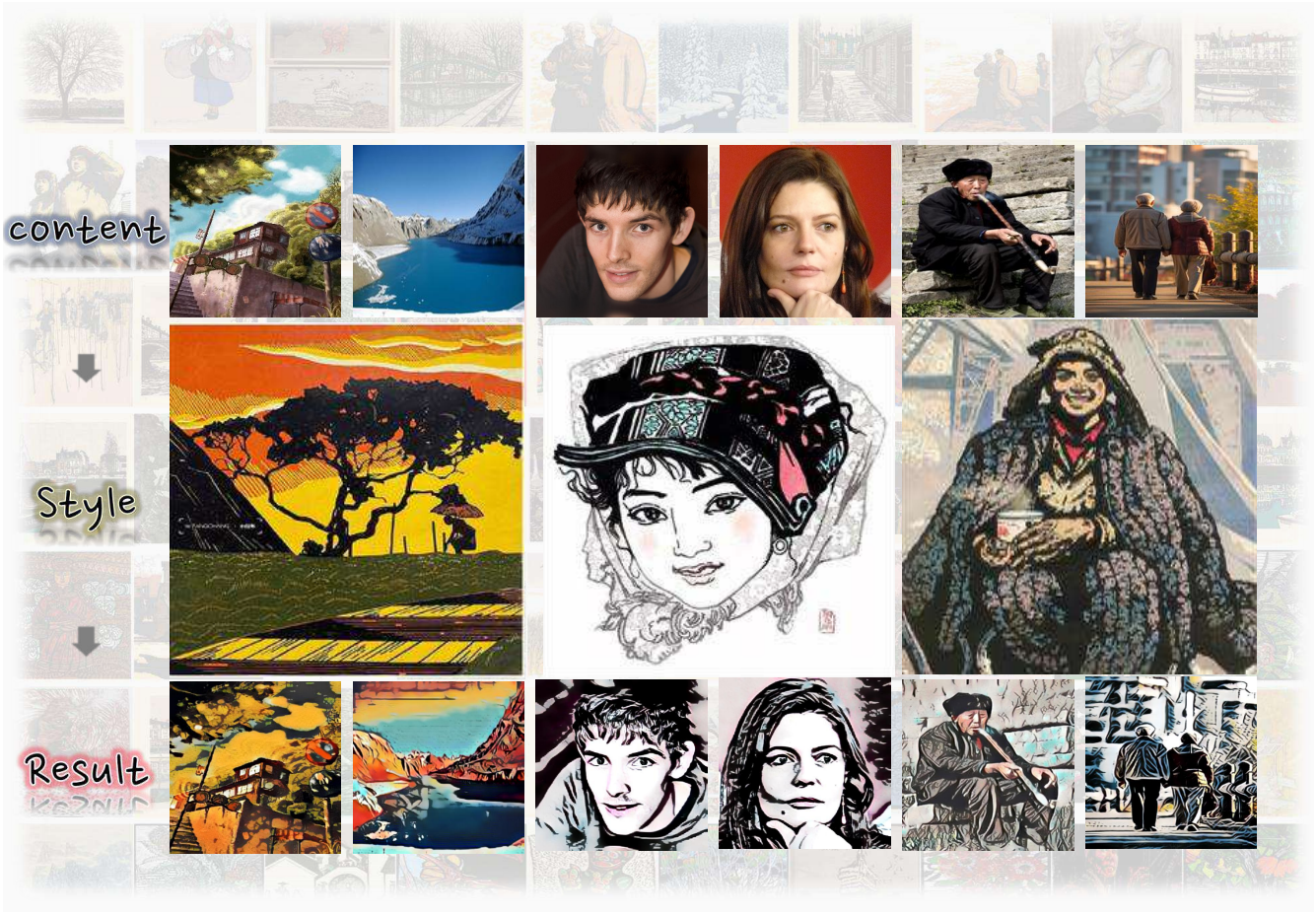


Figure 1. Visualization of style transfer effects for woodblock prints.

woodblock paintings. The main contributions of this work are as follows:

• We propose a Content-Aware Positional Encoder that combines spatial positional encoding with the Transformer architecture. This module enhances the extraction of global semantic features, leading to better structural preservation in the generated images. This addresses the issue of insufficient global feature learning and improves overall structural consistency.

• We introduce a multi-scale style encoder, termed the HierStyle Encoder, which captures hierarchical style features. These features are progressively fused with content representations through a decoder. This design improves color and texture coherence, effectively addressing the challenge of style inconsistency and enabling natural color transitions and texture blending.

• We design a ChromaNova Color Decoder to dynamically optimize color embeddings, enhancing the color-semantic matching between the generated image and the target style. Additionally, we formulate a multi-dimensional color loss function $\mathscr{L}_{MDC}$, which jointly optimizes color style from multiple aspects, including histogram distribution, color moment matching, and perceptual consistency. This comprehensive loss function better captures the unique color characteristics of woodblock painting by considering both color contrast and distribution differences.

We validate the effectiveness and robustness of our method through comparative experiments, user studies, ablation studies, and generalization tests. Experimental results show that our model effectively mitigates issues such as insufficient color contrast and style distortion in existing methods.

2 Related Work

2.1 Monochrome and Ink-Wash Painting Synthesis

Xu et al.[12] introduced a stroke-based rendering approach tailored for ink-style painting generation. Yang et al.[13] proposed a technique to produce Chinese ink wash artworks by leveraging reusable brushstroke patterns. Most existing studies primarily concentrate on monochrome ink painting, with relatively limited exploration of color-based artistic styles. Furthermore, sketch inputs exhibit distinct characteristics compared to grayscale images. Several models[14], while effective on grayscale data,

struggle to accurately apply color to line drawings.

However, these approaches are confined to monochrome expression and are ill-suited for traditional polychrome styles such as woodblock prints, which rely on symbolic color palettes and discrete chromatic regions.

2.2 Colorful Style Transfer and Line Drawing Colorization

To better capture both global color perception and local chromatic details, recent works have explored specialized strategies such as customized loss functions[15] and the integration of reference images to enhance the colorization process. DDcolor[16] introduces a Colorfulness loss that integrates image colorfulness to measure image quality. Meanwhile,[17] uses color histogram loss to measure the distribution distance between the output color histogram and the reference image, providing an intuitive way to describe image color. AesPA-Net[18] introduces Patch-wise Style Loss to convey the complex structure of local patterns, serving as a good constraint for coloring images with complex structural information.

Existing methods still lack a unified mechanism to jointly model structural semantics of line drawings and the stylistic color conventions inherent in traditional art forms.

2.3 Deep Architectures for Global-Local Style Modeling

In the domain of image style transfer[19], significant progress has been made in guiding artistic image generation. CAST[20] improves arbitrary style transfer by learning style representations through contrastive learning directly from image features. Beyond CNNs, Transformer-based architectures have also been applied to style transfer tasks. Wu et al.[21] proposed StyleFormer, a method that enables content-aware global style synthesis using a dedicated style synthesis module. Deng et al.[22] presented StyTr2, which mitigates content distortion during style transfer by modeling long-range dependencies within the input image. Despite their success in capturing overall style and tone, these methods often fall short in preserving fine-grained textures and hierarchical details.

In recent years, adaptive normalization techniques have gained popularity in style transfer. SANet[23] introduces a style attention mechanism that allows region-specific adjustments based on local style features. Liu et al.[24] further extended AdaIN[25] by proposing AdaAttN, which learns attention features at both pixel and feature levels from multi-level representations to achieve more realistic style blending.

When applied to high-contrast, discretely colored styles like woodblock prints, these architectures often fail to simultaneously preserve sharp structural boundaries and fine local chromatic details.

2.4 Learning Paradigms of Supervised, Unsupervised, and Semi-supervised

There are three training methods for style transfer. 1) Supervised methods: Image-to-image style transfer[26] has reached state-of-the-art results. 2) Unsupervised methods: Unsupervised image-to-image translation[27][28] is a typical solution for complex scene stylization without paired training data. 3) Semi-supervised methods: Oza et al.[29] address the instability of cyclic consistency constraints in challenging tasks by learning image transformations in the latent feature space rather than directly in the image space. Jiang et al.[30] propose generating anime scenes with a novel framework called Scenimefy, which leverages a small amount of paired data alongside a large quantity of unpaired data.

Given the scarcity of expert-annotated paired data in heritage art domains, there remains a critical need for a semi-supervised framework that effectively leverages limited labeled examples alongside abundant unpaired artistic imagery.

2.5 Loss Function

In image style transfer research, the loss function focuses on content and style constraints. Traditional pixel-level difference-based loss functions struggle to maintain color consistency and detail. DDcolor[16] introduces a Colorfulness loss that integrates image colorfulness to measure image quality. Meanwhile,[17] uses color histogram loss to measure the distribution distance between the output color histogram and the reference image, providing an intuitive way to describe image color. AesPA-Net[18] introduces Patch-wise Style Loss to convey the complex structure of local patterns, serving as a good constraint for coloring images with complex structural information. On the other hand, CAST[20] proposes a new contrast style loss based on global style distribution for transformation to accomplish arbitrary image style transfer.

3 Methodology

Multi-color woodblock painting contain richer RGB information, disordered color layouts, and abstract semantic contexts, which make feature extraction challenging for models. This often leads to issues such as semantic style distortion and low color contrast, significantly affecting the generation of realistic style textures. To address these challenges, we improve upon the StyTr2[22] model and propose a style transfer framework based on multi-scale color embedding. Our method aims to enhance the artistic expression and overall coherence of the generated images, while preserving the original color proportions and layout. Specifically, we employ a HierStyle Encoder to extract hierarchical style representations from the style image. These features,

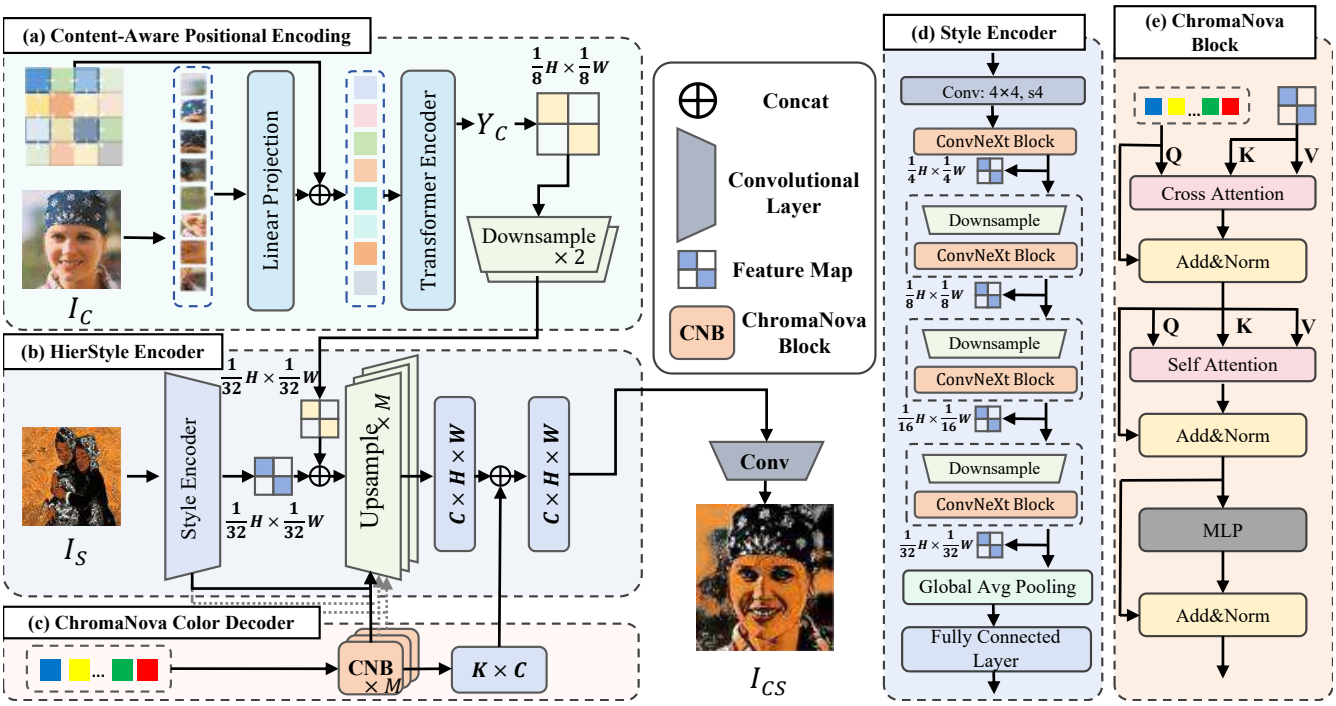


Figure 2. Overall framework. (a) **Content-Aware Positional Encoder**: Built upon a Vision Transformer architecture, it extracts global semantic features through dynamic patching and content-aware positional encoding. (b) **HierStyle Encoder**: Fuses content and style features to generate high-resolution image embeddings. (c) **ChromaNova Color Decoder**: Gradually refines color embeddings using multi-scale visual features, enabling semantic-aware color prediction. (d) **Style Encoder**: Extracts multi-scale style features using a backbone composed of downsampling layers and stacked Style Blocks. (e) **ChromaNova Block**: A lightweight module for color-semantic feature interaction.

together with the global semantic features extracted by the content encoder, are fed into a progressive decoder. The decoder performs multi-stage upsampling with cross-layer connections to reconstruct the image. In addition, we design a color decoder that receives multi-scale visual features to guide and refine the color embedding representation. Finally, a multi-dimensional color loss function is introduced to improve the ability to learn color characteristics. This helps alleviate the issues of low color contrast and semantic distortion in multi-color woodblock style transfer.

3.1 Overall Framework

We propose a HierStyle Multi-scale Semantic Style Transfer Method for Chinese Woodblock Painting, as illustrated in Fig. 2. The architecture consists of the following components: (a) **Content-Aware Positional Encoder (CAPE)**: A Vision Transformer-based module that extracts global semantic features from the content image using dynamic patching and content-aware positional encoding. (b) **Style Encoder**: A CNN-based encoder with downsampling layers and Style Blocks that extracts multi-scale style features from the style image. (c) **Progressive Decoder**: A multi-stage upsampling network that fuses content and style features via skip connections to reconstruct high-resolution embeddings. (d) **ChromaNova Color Decoder**: A Transformer-based decoder that refines learnable color queries through cross-attention with multi-scale visual features. (e) **ChromaNova Block**: A lightweight building unit within the color decoder that enables semantic-aware color prediction.

3.2 Content-Aware Positional Encoder

The Content-Aware Positional Encoder module extracts semantic structure features from the content image. It first divides the image into fixed-size patches, which are projected into embedding sequences through a linear layer, preserving spatial structure information. A multi-layer Transformer encoder is then used to capture global semantic dependencies. Each encoder layer consists of a multi-head self-attention module and a feedforward neural network, with residual connections and layer normalization applied to stabilize training.

Given a content image $I_c \in \mathbb{R}^{H\times W\times 3}$ and a multi-color woodblock style image $I_s \in \mathbb{R}^{H\times W\times 3}$, where H, W, and 3 denote the height, width, and number of channels respectively, the input images are first divided into patches of size $m \times m$ along the spatial dimensions.

Next, linear projection is applied to generate the embedded feature sequence $\varepsilon_c \in \mathbb{R}^{L\times C}$, where $L = \frac{H\times W}{m^2}$. Unlike conventional fixed-size patching, this module dynamically adjusts the number of patches to support robustness to different input resolutions. This patching strategy also helps decompose global image information into local features, facilitating subsequent feature extraction and processing.

To support variable input resolutions, the positional encoding is not fixed but adaptively generated based on content semantics (see Eq. 3), ensuring consistent spatial reasoning across scales.

To capture structural information, positional encoding is incorporated into the input sequence when using Transformer-based models. The attention score between the i-th and j-th patch is computed as shown in Eq. 1:

$$A_{i,j} = [(\varepsilon_i + P_i)W_q]^T[(\varepsilon_j + P_j)]W_k] \tag{1}$$

where W_q and W_k are learnable parameter matrices for query and key computation, and P_i denotes the 1D positional encoding of the i-th patch. In a 2D setting, the positional relationship between patches at (x_i, y_i) and (x_j, y_j) is calculated based on spatial distance:

$$P(x_i,y_i)^T P(x_j,y_j) = \sum_{k=0}^{\frac{d}{4}-1} [\cos(W_k(x_j - x_i)) + \cos(W_k(y_j - y_i))] \tag{2}$$

where $W_k = 1/10000^{2k/128}$ and $d = 512$. The positional relationship between patches depends only on their spatial distance.

Content-Aware Positional Encoding (CAPE) is scale-invariant, making it more suitable for style transfer tasks. Assuming an $n \times n$ positional encoding is sufficient to represent image semantics, for an input image $I_c \in \mathbb{R}^{H\times W\times 3}$, the fixed $n \times n$ encoding is resized to $\frac{H}{m} \times \frac{W}{m}$. This ensures that spatial relationships between patches remain consistent across different image scales:

$$P_L = F_{pos}(AvgPool_{n\times n}(\varepsilon_c)) \tag{3}$$

where $AvgPool_{n\times n}$ is an average pooling function, F_{pos} denotes a learnable positional encoding function implemented via 1x1 convolution, and P_L represents the content-aware positional encoding aligned with the sequence ε_c. In our experiments, $n = 18$.

The final CAPE for a patch at (x, y), denoted as $P_{CA}(x, y)$, is defined as:

$$P_{CA}(x,y) = \sum_{k=0}^{s}\sum_{l=0}^{s}(a_{kl}P_L(x_k, y_l)) \tag{4}$$

where a_{kl} is the interpolation weight and s denotes the number of neighboring patches. The final feature embedding of a patch is obtained by adding its input embedding ε_c and the corresponding P_{CA}.

After this, we obtain the content sequence:

$$\mathbb{Z}_c = \{S_{c1} + P_{cA1}, S_{c2} + P_{cA2}, ..., S_{cL} + P_{cAL}\} \tag{5}$$

which is fed into the Transformer encoder. Each layer of the encoder consists of a multi-head self-attention module (MSA) and a feed-forward network (FFN). The input sequence $\mathbb{Z}_c$ is projected into query (Q), key (K), and value (V) vectors, where $W_q, W_k, W_v \in \mathbb{R}^{c\times d_{head}}$.

$$Q = Z_c W_q, K = Z_c W_k, V = Z_c W_v \tag{6}$$

$$F_{MSA}(Q,K,V) = Concat(Att_1(Q,K,V), \quad \ldots, Att_N(Q,K,V))W_0 \tag{7}$$

where $W_0 \in \mathbb{R}^{C\times C}$ is a learnable parameter, N is the number of attention heads, and $d_{head} = \frac{C}{N}$ is the dimension of each head. The outputs of all heads are concatenated and projected via W_0. The attention computation for the i-th head is defined as:

$$Att_i(Q,K,V) = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_{\text{head}}}}\right) V_i \tag{8}$$

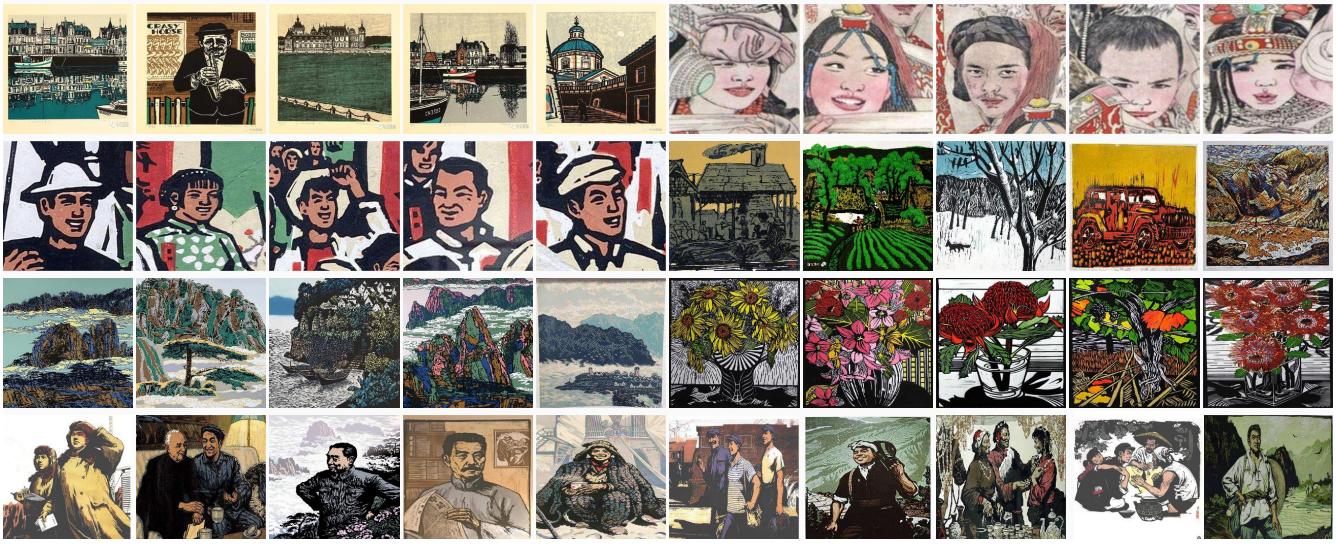


Figure 3. Visualization display of WoodCut dataset.

Residual connections are applied to stabilize training:

$$Y_c^{'} = F_{MSA}(Q,K,V) + Q, \quad Y_c = F_{FFN}(Y_c^{'}) + Y_c^{'} \tag{9}$$

where $F_{FFN}(Y_c^{'}) = max(0, Y_c^{'}W_1 + b_1)W_2 + b_2$. Layer normalization (LN) is applied after each module. After passing through multiple Transformer encoder layers, the final content feature sequence is obtained as $Y_c \in \mathbb{R}^{L\times C}$.

To meet the input requirements of the progressive decoder, the content features are concatenated with the style features. To ensure spatial and channel compatibility, the Transformer output Y_c is reshaped into a spatial feature map and downsampled twice to match the spatial size of the style feature map. The resulting content feature map is denoted as $F_c \in \left[B, C, \frac{H}{32}, \frac{W}{32}\right]$, which preserves both spatial and global semantic information from the original content image.

3.3 HierStyle Encoder

Given a style image $I_s \in \mathbb{R}^{H\times W\times 3}$, the HierStyle Encoder extracts high-level style information to enhance semantic and texture coherence. The encoder backbone consists of stacked downsampling layers and Style Blocks, which are chosen for their strong multi-scale feature extraction capability. This design enables the model to learn hierarchical image representations that are crucial for style-aware color prediction.

The network produces four intermediate feature maps: $F_{s1}, F_{s2}, F_{s3}, F_{s4}$, with spatial resolutions of $\frac{H}{4}\times\frac{W}{4}$, $\frac{H}{8}\times\frac{W}{8}$, $\frac{H}{16}\times\frac{W}{16}$, and $\frac{H}{32}\times\frac{W}{32}$, respectively. The first three feature maps are passed to the progressive decoder via skip connections.

Before fusion, a 1×1 convolution is applied to adjust the channel dimension of the content feature map F_c to match that of the highest-level style feature map F_{s4}. These two feature maps are then concatenated along the channel dimension and used as the main input to the progressive decoder.

The decoder consists of four stages that progressively increase the image resolution. Each stage includes an upsampling layer and a shortcut layer. Unlike conventional methods that use deconvolution or interpolation, we employ Pixel Shuffle for upsampling. This layer rearranges sub-pixels to avoid checkerboard artifacts and improve reconstruction quality. It transforms a low-resolution feature map of shape $\left(\frac{h}{p}, \frac{w}{p}, cp^2\right)$ into a high-resolution one of shape (h, w, c).

The shortcut layers use convolution and skip connections to integrate features from the corresponding stages of the style encoder. Through this progressive upsampling process, the decoder captures the full image feature pyramid. These multi-scale features are further used as input to the color decoder to guide color embedding optimization. The final output of the progressive decoder is an image embedding $E_{cs} \in \mathbb{R}^{C\times H\times W}$, with the same spatial resolution as the input image. By fusing multi-stage style features and global semantic content information, this architecture provides a high-quality feature representation that is both semantically accurate and style-consistent across scales.

The encoder outputs four feature maps at decreasing resolutions, which are later fused to guide hierarchical stylization.

The multi-scale feature hierarchy not only captures style at different granularities but also enhances interpretability: high-level features govern global attributes such as dominant color palette and composition, while low-level features control fine

details like texture and stroke patterns—aligning with human perception of artistic style.

3.4 ChromaNova Color Decoder

The ChromaNova Color Decoder is constructed by stacking multiple ChromaNova Blocks, each receiving visual features and color queries as inputs. Based on an enhanced Transformer decoder architecture, the CNB's design is illustrated in Fig. 2(e).

To learn adaptive color queries based on visual semantic information, we introduce a learnable color embedding memory for storing a sequence of color representations: $Z_0 = \left[Z_0^{(1)}, Z_0^{(2)}, \ldots, Z_0^{(k)}\right] \in \mathbb{R}^{K\times C}$, where K denotes the number of embeddings, and C represents the channel dimension. These color embeddings are initialized to zero at the beginning of training and used as initial color queries in the first CNB.

Semantic representation-color embedding correlation is established through a cross-attention layer, as shown in Eq. 10:

$$Z'_l = \text{softmax}\left(\frac{Q_l K_l^T}{\sqrt{C}}\right) V_l + Z_{l-1}, \tag{10}$$

where l is the network layer index, $Z_l \in \mathbb{R}^{K\times C}$ represents the color embedding vector at layer l, $Q_l = f_Q(Z_{l-1}) \in \mathbb{R}^{K\times C}$ is the query matrix, and $K_l, V_l \in \mathbb{R}^{H_l\times W_l\times C}$ are transformed image features processed by functions $f_K(\cdot)$ and $f_V(\cdot)$ respectively.

Following the cross-attention operation, standard Transformer layers further process the color embeddings as defined in Eq. 11:

$$\begin{aligned} Z''_l &= \text{MSA}\left(\text{LN}(Z'_l)\right) + Z'_l \\ Z'''_l &= \text{MLP}\left(\text{LN}(Z''_l)\right) + Z''_l \\ Z_l &= \text{LN}(Z'''_l) \end{aligned} \tag{11}$$

where $\text{MSA}(\cdot)$ stands for multi-head self-attention, $\text{MLP}(\cdot)$ for multilayer perceptron, and $\text{LN}(\cdot)$ for layer normalization. In our proposed CDB module, the cross-attention operation precedes the self-attention operation to enrich color embeddings with semantic information before applying self-attention.

To balance computational complexity and representational power, three image features at different scales are utilized. Specifically, intermediate visual features with down-sampling rates of 1/16, 1/8, and 1/4 from the progressive decoder are used in the color decoder. These modules are grouped into sets of three, where multi-scale features are sequentially fed into each CDB. The entire group is executed cyclically M times, resulting in a total of 3M CDBs in the color decoder. This process can be formalized as:

$$E_c = \delta(Z_0, F_1, F_2, F_3) \tag{12}$$

where F_1, F_2, F_3 represent visual features at three different scales.

Employing multi-scale features in the color decoder allows modeling the relationship between color queries and visual embeddings, progressively associating semantic regions and producing dynamic color embeddings. This ensures that the color embeddings $E_{color} \in \mathbb{R}^{K\times C}$ are more sensitive to semantic boundaries, enhancing the accuracy of semantic boundary identification and ensuring color consistency.

Each color query interacts with multi-scale content features via cross-attention to produce spatially adaptive color responses. Each color query $z^{(k)}$ acts as a learnable color concept, and its cross-attention weights over spatial locations indicate where in the content image this color should be applied. This attention-based assignment provides an intrinsic mechanism for interpreting how style colors are semantically mapped onto content structures.

3.5 Feature Fusion

The feature fusion module aims to integrate outputs from the Content-Aware Positional Encoder and ChromaNova, producing a colorized Woodblock painting style image. It takes two primary inputs: image embeddings $E_{cs} \in \mathbb{R}^{C\times H\times W}$ from the HierStyle Encoder and color embeddings $E_{color} \in \mathbb{R}^{K\times C}$ from the ChromaNova Color Decoder, where C is the embedding dimension and K represents the number of color queries.

Initially, a semantic association between image embeddings and color queries is established through matrix multiplication. A dot product operation aggregates these embeddings to produce an affinity weight matrix $\widehat{\mathscr{F}} \in \mathbb{R}^{K\times H\times W}$. The spatial features are then reshaped to adjust the dimensions of the weight matrix $\widehat{\mathscr{F}}$, restoring the spatial structure. At this stage, each spatial location

(i, j) has a feature vector $\widehat{\mathscr{F}}_{i,j} \in \mathbb{R}^K$ that indicates the association strength with K color queries. To map the K-dimensional stylistic-color features to RGB space, a 1×1 convolution layer is employed for channel compression as shown in Eq. 13:

$$I_{cs} = Conv_{1\times1}(\widehat{\mathscr{F}}) \tag{13}$$

where the convolution kernel parameters $W_{rgb} \in \mathbb{R}^{3\times K\times1\times1}$ dynamically map the K channels to three RGB channels, generating the final stylized output $I_{cs} \in \mathbb{R}^{H\times W\times3}$.

3.6 Loss Function

The overall training objective combines a proposed Multi-Dimensional Color Loss ($\mathscr{L}_{\text{MDC}}$) with standard content and style losses via weighted summation.

Multi-Dimensional Color Loss. To capture the high-contrast and hue-concentrated characteristics of woodblock prints, we design a differentiable and perceptually motivated color loss:

$$\mathscr{L}_{\text{MDC}} = \sum_{c\in\{a,b\}} \left(\left| \mu^c_{\text{gen}} - \mu^c_{\text{style}} \right| + \left| \sigma^c_{\text{gen}} - \sigma^c_{\text{style}} \right| \right) \quad + \left\| \mathbf{G}(F_{\text{gen}}) - \mathbf{G}(F_{\text{style}}) \right\|_F^2, \tag{14}$$

where c denotes the chromatic channels in CIELAB space (a: green–red, b: blue–yellow), and μ^c, σ^c are the channel-wise mean and variance computed over spatial dimensions. The first term enforces statistical similarity in color concentration and dispersion, which is fully differentiable and directly controls global hue balance.

For the second term, we define the color Gram matrix as $\mathbf{G}(F) = \frac{1}{HW} F^\top F$, where $F \in \mathbb{R}^{HW\times C}$ is the feature map reshaped from the output of a pre-trained CNN (VGG-19, relu3_1) fed with the luminance channel L and chrominance noise. This Gram-based term captures second-order color correlations across spatial locations, promoting smooth transitions and perceptual consistency—key for mimicking pigment diffusion in woodblock prints. Both terms are smooth and differentiable, enabling end-to-end optimization.

Content and Style Losses. We adopt standard VGG-based losses:

$$\mathscr{L}_c = \frac{1}{N_l} \sum_{i=1}^{N_l} \left\| \phi_i(I_{cs}) - \phi_i(I_c) \right\|_2, \tag{15}$$

$$\mathscr{L}_s = \frac{1}{N_l} \sum_{i=1}^{N_l} \Big(\left\| \mu(\phi_i(I_{cs})) - \mu(\phi_i(I_s)) \right\|_2 + \left\| \sigma(\phi_i(I_{cs})) - \sigma(\phi_i(I_s)) \right\|_2 \Big), \tag{16}$$

where $\phi_i(\cdot)$ denotes features from the i-th layer of VGG-19, and $\mu(\cdot)$, $\sigma(\cdot)$ are spatial mean and standard deviation.

Joint Loss. The total loss is:

$$\mathscr{L}_{\text{Joint}} = \lambda_1 \mathscr{L}_{\text{MDC}} + \lambda_2 \mathscr{L}_c + \lambda_3 \mathscr{L}_s, \tag{17}$$

with $\lambda_1 = 1.0$, $\lambda_2 = 0.6$, $\lambda_3 = 0.9$.

4 Experiments

4.1 Datasets

We introduce a woodblock painting dataset, termed **WoodCut**. In our experiments, we use the **WoodCut** dataset as the style images, and select an equal number of images from the **CelebA** dataset as content images, with 2000 images for training and 100 images for testing. Additionally, to evaluate the generalization ability of our method, we conduct tests on the anime dataset **AnimeGAN**, the natural landscape dataset **Places**, and the Chinese landscape paingting dataset **Landscape**.

• **WoodCut.** To deeply explore and preserve traditional Chinese Woodblock painting art, we construct WoodCut, a dedicated dataset for multi-color woodblock prints, as shown in Fig. 3. WoodCut covers a wide range of styles and historical periods, ensuring diversity. All collected images undergo a rigorous manual review process to guarantee quality and relevance. During this process, each image is carefully inspected and cropped, and samples that fail to meet quality standards are removed. Finally, by incorporating works from multiple authoritative institutions and renowned artists, WoodCut contains 1,200 high-resolution multi-color Woodblock painting images, representing diverse styles across modern and contemporary periods, as well as regional and temporal variations.

To meet algorithm training requirements, we preprocess the images in the dataset by resizing and format normalization, ensuring effective use in model training. For performance evaluation, we split the dataset into a training set of 1,100 images



Figure 4. Illustration of the Datasets. (a) is the WoodCut dataset, (b) is the CelebA dataset, (c) is the Landscape dataset, (d) is the AnimeGAN dataset, (e) is the Places dataset.

and a test set of 100 images. This split enables reliable validation of model effectiveness on tasks such as style transfer and image generation.

• **CelebA**[31]**.** As shown in Fig. 4(b), As a commonly used image coloring dataset, CelebA contains a series of celebrity face images that can be rendered more vivid and lifelike by coloring gray-scale face images.

• **Landscape.** As shown in Fig. 4(c), W collect images from a collaborative agency and Digital Art Database Collection. Landscape comprises 2,969 landscape paintings from various dynasties, featuring a wide range of color styles. During the construction of the dataset, we allocated 2,769 images for the training set, and 100 images each for the validation and test sets. Through further random crop and processing, we ultimately formed a training set containing 10,428 images, along with validation and test sets comprising 800 images each.

• **AnimeGAN**[32]**.** As illustrated in Fig. 4(d), the Anime-style image dataset was constructed by collecting 143,000 anime images from the website danbooru.donmai.us using the gallery-dl crawler tool. Subsequently, an anime detection script, python-anime, was employed to filter and retain only those images that conform to the anime style.

• **Places**[33]**.** As shown in Fig. 4(e), the Places dataset is a large-scale scene recognition database consisting of over 1.8 million natural images covering 365 distinct scene categories. It is widely used for evaluating models in tasks such as scene understanding, domain adaptation, and image synthesis. In our experiments, we follow the standard evaluation protocol and select the same 800 images from the validation set for consistent performance comparison with existing methods.

4.2 Pre-processing

We employed a method of random uniform sampling to obtain 10,000 samples with a resolution of 256×256 from the original dataset for model training. To ensure the reliability of the sampled features, we employed an aesthetic evaluation metrics algorithm to screen the collected images. This process retains data with pronounced landscape painting attributes while filtering out irrelevant or extraneous content. By doing so, it guarantees both the consistency and artistic quality of the dataset.

4.3 Implementation Details

All experiments are conducted on an NVIDIA GeForce RTX 4090 GPU with 24GB memory. The model is implemented and trained using PyTorch 2.1 and Python 3.10. We optimize the model using the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$) with a linearly decaying learning rate. The initial learning rate is set to 5×10^{-4}, and the batch size is 4. Training runs for a total of 160,000 iterations until convergence.

The proposed **HierWoodNet** consists of three main components: (1) A *Content-Aware Positional Encoder* built upon a Vision Transformer (ViT) backbone with 12 layers, hidden dimension of 768, and 12 attention heads; (2) A *HierStyle Encoder* that employs a U-Net-like architecture with four downsampling levels (feature dimensions: $64 \rightarrow 128 \rightarrow 256 \rightarrow 512$), each level incorporating residual blocks and multi-scale style modulation; (3) A *ChromaNova Color Decoder* with three upsampling stages, where color embeddings are dynamically refined through channel-wise attention and hue-aware normalization.

We will release the source code, trained models, and dataset splits upon publication to ensure full reproducibility.



Figure 5. Visual comparison of migration effects with SOTAs.

4.4 Evaluation Metrics

We select standard metrics including LPIPS[34], FID[35], PSNR[36], SSIM[37], NIMA[38], and CLIP Score[39].

LPIPS and SSIM were used to assess the perceived similarity and structural similarity between the generated image and the real image, while PSNR was used to evaluate the accuracy of reconstructing landscape painting content images, and FID was used to measure the difference in distribution between generated and real images. In addition, NIMA (Neural Image Assessment) provides a perceptual estimate of aesthetic quality by predicting mean human opinion scores, and the CLIP Score quantifies semantic alignment between the generated image and the target woodblock style through cosine similarity in the CLIP multimodal embedding space.

5 Main Result

5.1 Baselines

To illustrate the effectiveness of our model, we choose four baselines as follows:

• **CAST[20].** It addresses arbitrary image style transfer by learning style representations directly from image features rather than second-order statistics.

• **AdaAttN[24].** It introduces Adaptive Attention Normalization, a novel module designed to improve arbitrary neural style transfer by adaptively performing attentive normalization on a per-point basis.

• **StyTr2[22].** It presents a transformer-based method for image style transfer, addressing the challenge of biased content representation by leveraging long-range dependencies in input images.

• **SANet[23].** It introduces a novel style-attentional network that effectively integrates local style patterns based on the semantic spatial distribution of the content image.

• **StyleDiffusion[40].** It achieves content-style disentanglement in the CLIP space using diffusion models, combining a CLIP-based disentanglement loss with style reconstruction priors for interpretable and controllable style transfer.

• **ContextDiff[41].** It enhances cross-modal alignment in diffusion models by incorporating text-visual context into both forward and reverse processes, enabling more semantically consistent text-guided generation.

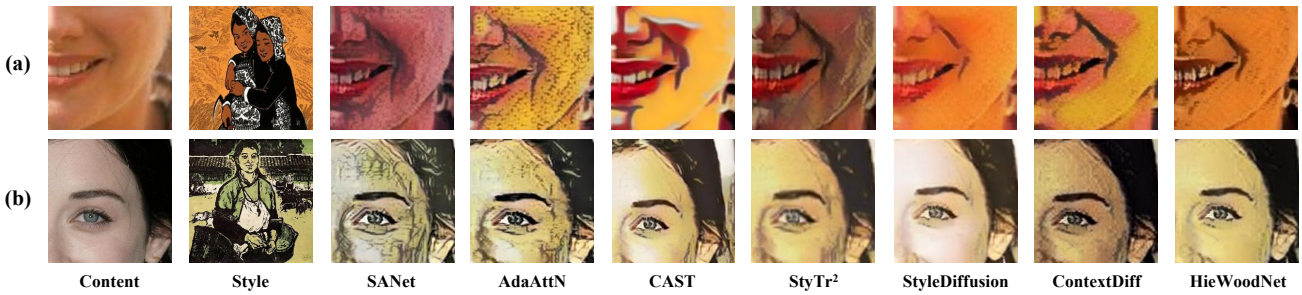


Figure 6. Visual comparison of transfer effect and SOTA details.

5.2 Comparison with State-of-the-Art Models

5.2.1 Comparison Analysis

Table 1. Comparison of SoTA methods. ↑ Higher values are better, ↓ Lower values are better. Optimal results are displayed in **bold**, while suboptimal results are underlined.

Method	SSIM↑	PSNR↑	FID↓	LPIPS↓	NIMA↑	CLIP↑
CAST[20]	0.849	25.416	121.416	0.271	4.77	0.2157
StyTr²[22]	0.856	26.133	120.467	0.250	4.83	0.2113
AdaAttN[24]	0.825	23.092	125.902	0.268	5.16	0.1983
SANet[23]	0.817	22.783	124.433	0.284	5.33	0.2022
StyleDiffusion[40]	**0.882**	25.417	121.102	0.253	5.49	**0.2256**
ContextDiff[41]	0.853	26.593	119.734	0.244	5.28	0.2145
Ours	0.873	**27.101**	**118.032**	**0.241**	**5.54**	0.2238

To evaluate the effectiveness of our method on color woodblock painting style transfer, we compare against several state-of-the-art approaches, including CAST[20], StyTr²[22], SANet[23], AdaAttN[24], as well as recent diffusion-based methods StyleDiffusion[40] and ContextDiff[41].

As shown in Table 1, our method achieves the best performance in PSNR, FID, and LPIPS, and ranks second in SSIM—demonstrating superior content preservation, visual fidelity, and perceptual quality. While StyleDiffusion excels in structural similarity, it falls short in color contrast and semantic alignment for woodblock styles. ContextDiff improves cross-modal consistency but lacks specialized modeling for traditional art characteristics such as edge diffusion and hue concentration. In contrast, our HierStyle encoder captures hierarchical style cues, the progressive decoder ensures coherent structure–style fusion, and the ChromaNova color decoder—guided by the multi-dimensional color loss effectively reproduces the high-contrast, symbolic color semantics unique to woodblock prints.

5.2.2 Visualization

Fig. 5 presents a visual comparison of stylized results across several state-of-the-art methods. SANet produces outputs with disordered textures and weak style coherence. ContextDiff improves color–semantic alignment but often yields unnatural or fragmented color distributions. StyleDiffusion achieves smoother color transitions yet suffers from semantic inconsistency, particularly in complex structural regions. CAST preserves color fidelity well but tends to generate visually monotonous results with limited stylistic richness. More recent diffusion-based approaches, such as StyleDiffusion and ContextDiff, better capture global style semantics; however, StyleDiffusion sometimes oversmooths fine details, while ContextDiff struggles to maintain accurate color proportions in multi-color woodblock painting scenarios. In contrast, our method consistently generates results that are both visually appealing and faithful to the intricate color layout and artistic characteristics of traditional Chinese woodblock prints, demonstrating superior style expressiveness and semantic integrity.

Fig. 6 visualizes the fine-grained transfer quality of facial regions, highlighting key differences in style fidelity and detail preservation. StyleDiffusion produces smooth transitions but often deviates from the authentic color palette of traditional woodblock prints. ContextDiff better preserves color tones yet yields harsh or unnatural boundaries due to rigid local matching. In contrast, our method achieves both accurate color reproduction and natural texture blending, closely adhering to the stylistic essence and delicate chromatic characteristics of Chinese woodblock art.

5.3 User Study

To further validate the effectiveness, we conducted two user studies to evaluate our model against mainstream style transfer models through subjective scoring.

5.3.1 Style Expressiveness

We recruited 30 participants who were shown 50 randomly selected content-style image pairs for each method in a randomized order. Participants rated images based on three criteria: (a) conformity to color Woodblock painting style, (b) preservation of background information from the content image, and (c) overall style consistency with the style image. The rating scale was from 1 to 5. As illustrated in Fig. 7, our model outperforms others in learning the style and color characteristics of traditional Chinese landscape paintings while preserving sketch content more accurately. Our coloring effects also meet human aesthetic preferences, demonstrating superior generalization ability.

A paired t-test (two-tailed, $\alpha = 0.05$) confirms that the improvements over the best baseline are statistically significant across all three criteria ($p < 0.01$).

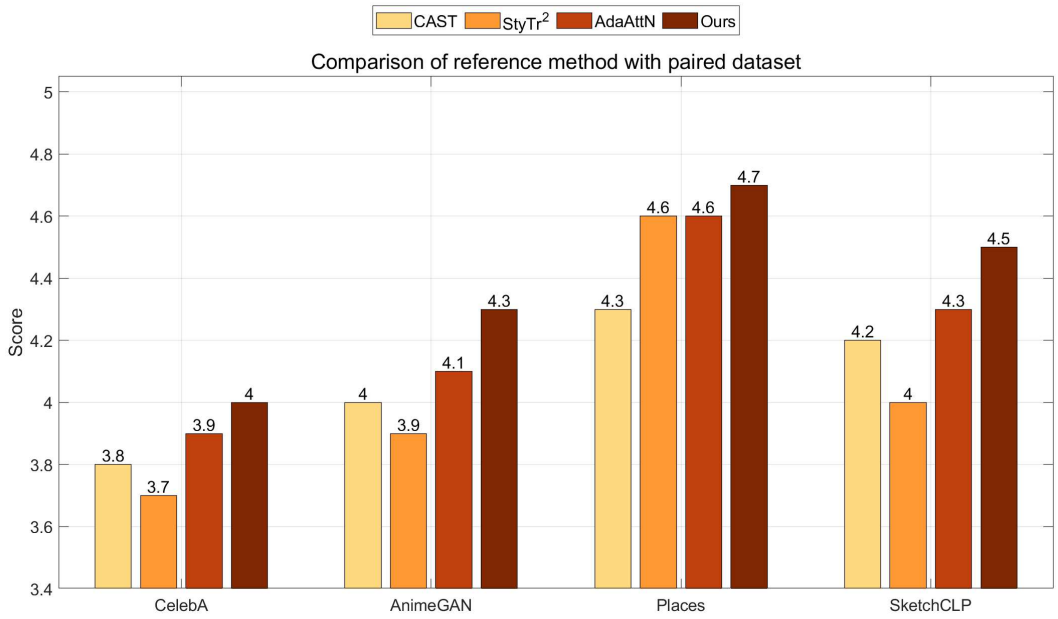


Figure 7. User study with state-of-the-art methods.

5.3.2 Style Authenticity Test

Thirty participants assessed the authenticity of stylized images. Each participant was shown 10 artworks per session, including 2 to 4 stylized images, and asked to identify synthetic ones. A total of 900 responses were collected. Accuracy (Acc) and Recall were used as metrics to measure the likelihood of recognizing synthesized images, as presented in Table 2. Results show that images generated by ours model were least likely to be identified as fake color woodblock painting.The difference in detection accuracy between our method and the strongest baseline is statistically significant (McNemar's test, $p = 0.003$).

Table 2. User study results for style authenticity test. Higher values are better unless marked ↓. Optimal results are in **bold**, suboptimal in underlined.

Method	Acc↑	Recall↑
CAST[20]	86.32%	85.43%
StyTr²[22]	**90.54%**	88.14%
AdaAttN[24]	80.90%	78.26%
SANet[23]	84.44%	83.51%
Ours	89.81%	**88.39%**



Figure 8. Visualization results of ablation experiments for each module.

5.4 Ablation Study

We performed ablation experiments on key components of model, including the ChromaNova Color Decoder, the multi-scale style feature extraction module HierStyle Encoder, and the Multi-Dimensional Color Loss. Additional ablation studies were designed for the HierStyle Encoder to verify its efficacy.

5.4.1 Effectiveness of Components

We conducted quantitative experiments by removing different modules to assess their impact on style transfer results, as shown in Table 3.

Comparison with w/o HierStyle: Removing this module decreases the SSIM by 4.9%, indicating that multi-scale style feature extraction significantly improves the adaptability to various color Woodblock painting styles. The NIMA score also drops from 5.54 to 5.32 and the CLIP score declines from 0.2238 to 0.2079, confirming reduced visual appeal and weaker alignment with the target style semantics. **Comparison with w/o ChromaNova**: The absence of this color decoder reduces PSNR by 21.8% and SSIM by 2.8%. This highlights its critical role in enhancing color consistency and optimizing style coloring. Moreover, both NIMA, 4.99 and CLIP, 0.1917 scores reach their lowest among all variants, demonstrating severe degradation in aesthetic quality and style fidelity. **Comparison with w/o MDC Loss**: Omitting this loss function lowers the PSNR by 14.5%, showing its importance in controlling global chroma distribution and enhancing color contrast. The corresponding NIMA,5.44 and CLIP,0.2106 scores are suboptimal, further verifying that MDC Loss contributes meaningfully to perceptual quality and semantic coherence.

Fig. 8 visualizes the results after removing different components. The model w/o HierStyle generates images with missing style details. The model w/o ChromaNova produces images with monotonous colors and lacking multi-level color transitions from the style image. The model w/o MDC Loss suffers from low color contrast. In contrast, the full model achieves superior results with strong color semantic consistency, high contrast, and well-balanced color distribution.

5.4.2 Effectiveness of HierStyle Encoder

To evaluate the effectiveness of the HierStyle Encoder, we replace it with ResNet-50[42], VGG-19[43], and a Transformer encoder, respectively. The results are summarized in Table 4.

ResNet-50 and VGG-19 show worse FID scores due to limited feature representation. VGG-19, in particular, tends to

Table 3. Ablation comparison on Components. ↑ Higher values are better, ↓ Lower values are better. Optimal results are displayed in **bold**, while suboptimal results are underlined.

Variants	SSIM↑	PSNR↑	FID↓	LPIPS↓	NIMA↑	CLIP↑
w/o HierStyle	0.830	24.416	183.611	0.273	5.32	0.2079
w/o ChromaNova	0.849	21.182	162.402	0.291	4.99	0.1917
w/o MDC Loss	0.865	23.184	159.930	0.287	5.44	0.2106
Ours	**0.873**	**27.101**	**118.032**	**0.241**	**5.54**	**0.2238**

Table 4. Comparison on HierStyle Encoder. ↑ Higher values are better, ↓ Lower values are better. Optimal results are displayed in **bold**, while suboptimal results are underlined.

Encoder	SSIM↑	PSNR↑	FID↓	LPIPS↓	Time(ms)↓	NIMA↑	CLIP↑
ResNet-50[42]	0.849	24.782	131.561	0.268	15.982	4.93	0.1966
VGG-19[43]	0.814	22.306	128.074	0.271	29.313	4.72	0.1835
Transformer[44]	**0.882**	26.891	122.953	0.253	48.745	5.21	0.2086
ConvNeXt(Ours)	0.873	**27.101**	**118.032**	**0.241**	**11.068**	**5.54**	**0.2238**

over-abstract texture features, leading to lower SSIM and PSNR. Both the ConvNeXt Block and Transformer encoder achieve better FID scores, indicating generated images are closer to real woodblock painting. However, as noted in prior work, the Transformer typically performs style attention on single-scale feature maps, which limits its ability to capture low-level semantic cues. This may cause color leakage in complex scenes. Among the backbones, ConvNeXt achieves the fastest style feature extraction, reducing inference time by over 30% compared to other models. Importantly, our ConvNeXt-based encoder yields the highest NIMA, 5.54 and CLIP, 0.2238, surpassing even the Transformer variant in perceptual quality and style-text alignment, which underscores its superiority in modeling hierarchical stylistic attributes for color woodblock painting synthesis.

5.5 Comparison on Diverse Datasets

To evaluate the generalization of our model, we conducted cross-dataset tests on AnimeGAN, Places, and Landscape. The style dataset is always WoodCut. The experimental results are summarized in Table 5, and visual comparisons are shown in Fig. 9.

AnimeGAN[32]: The performance metrics on AnimeGAN are close to those on CelebA, with a 1.24% decrease in LPIPS, indicating superior visual quality. As illustrated in Fig. 9(a) and (b), the style transfer for anime datasets achieves good overall results.

Places[33]: On the Places dataset, the PSNR increased by 3.90%, suggesting higher fidelity in image details compared to real images. However, $\mathscr{L}_c$ rose by 12.12%, indicating a slight content discrepancy between generated and real images. Fig. 9(c) and (d) show that the generated images largely preserve structural integrity with minimal style loss, achieving effective style transfer.

Landscape: Performance metrics on Landscape are also similar to those on CelebA. Specifically, FID decreased by 0.47%, implying better distribution alignment of generated images. Nevertheless, $\mathscr{L}_c$ increased by 12.48%, pointing to more significant content distortion at higher semantic levels. Fig. 9(e) and (f) demonstrate that Landscape can still

These findings confirm the robustness and adaptability of our model across diverse datasets.

6 Limitations

Although the proposed method achieves promising results in color Woodblock painting style transfer, there are still several limitations that suggest directions for future improvement.

Ours model effectively addresses key challenges such as color semantic distortion and texture disorder. To further enhance its performance, we plan to construct and annotate a more comprehensive Woodblock painting dataset with cultural and historical significance. This will improve the model's generalization across diverse artistic styles. In addition, we will explore model compression and quantization techniques to reduce computational cost. These optimizations aim to enable efficient deployment on mobile or interactive platforms, making the method more suitable for real-time artistic creation. Together, these efforts will contribute to the digital preservation of traditional woodblock painting and support the innovative inheritance and dissemination of intangible cultural heritage.

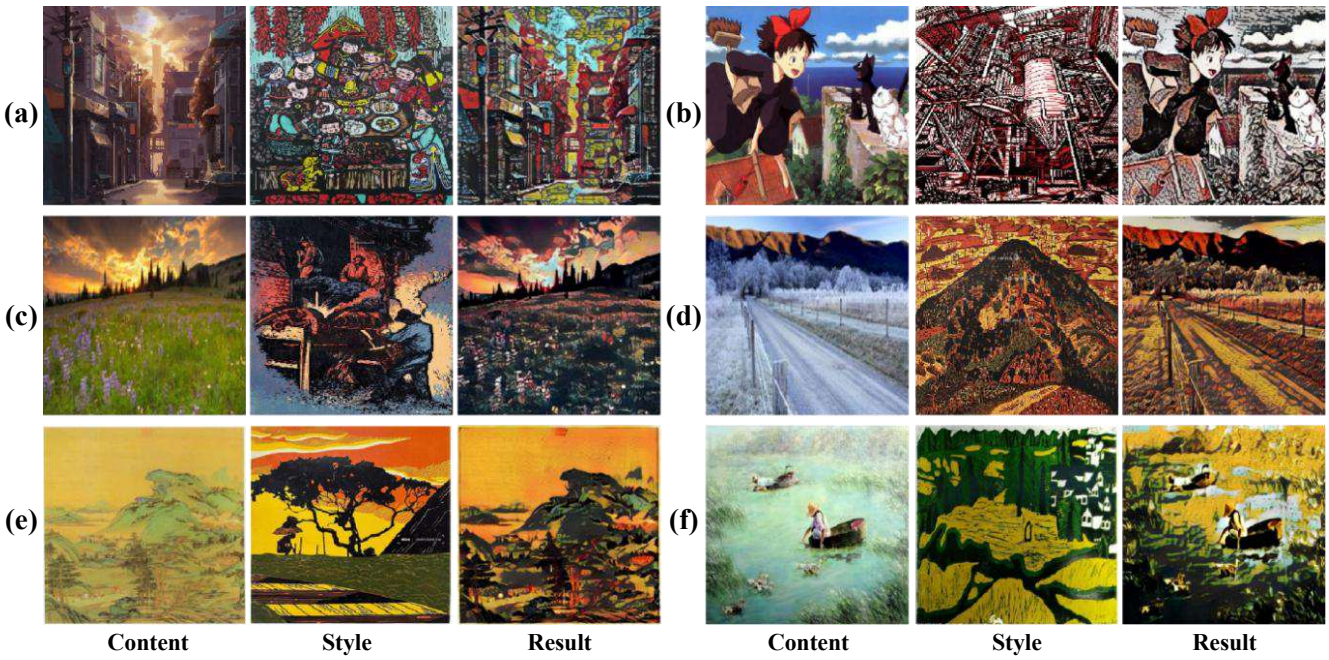


Figure 9. Visual comparison on diverse datasets. (a) and (b) Experiments on the AnimeGAN dataset, (c) and (d) Experiments on the Places dataset, (e) and (f) Experiments on the Landscape dataset.

Table 5. Comparison on diverse datasets. ↑ Higher values are better, ↓ Lower values are better. *Optimal results are displayed in **bold**, while suboptimal results are underlined.

Datasets	SSIM↑	PSNR↑	FID↓	LPIPS↓	NIMA↑	CLIP↑
AnimeGAN[32]	0.864	26.963	118.302	**0.238**	5.39	0.2173
Places[33]	0.859	**28.159**	121.096	0.252	5.22	0.2241
Landscape	0.863	25.045	**117.477**	0.247	5.47	0.2154
CelebA[31]	**0.873**	27.101	118.032	0.241	**5.54**	**0.2238**

7 Conclusion

Due to the rich color information, abstract semantic structure, high color contrast, and highly stylized lines and color blocks in multi-color woodblock prints, mainstream style transfer models often suffer from color semantic distortion, insufficient contrast, disordered texture distribution, and blurred color boundaries. To address these issues, we propose HierStyle, a multi-scale semantic framework for Chinese woodblock painting style transfer, designed to more accurately capture and transfer their unique visual language.

Our method employs a pre-trained network based on ConvNeXt blocks in the HierStyle Encoder to extract style color and texture features at multiple scales, enhancing the correlation between contextual semantics and local structures. A skip-connection and progressive decoder jointly guide the color decoder to optimize color representation. The proposed color decoder introduces a multi-scale color query mechanism that learns hierarchical color representations from different feature levels. Additionally, a multi-dimensional color loss function is designed to further improve chroma, contrast, and regional distinctness.

Author contributions statement

Jie Qiao: Preparation, resources, Methodology;

Fangqing Fan: Data curation, validation, Methodology;

References

1. Tsai, F. W. & Van der Reyden, D. Analysis of modern chinese paper and treatment of a chinese woodblock print. *The Pap. Conserv.* **21**, 48–62 (1997).
2. Barrett, T. H. Woodblock dyeing and printing technology in china, c. 700 ad: the innovations of ms. liu, and other evidence. *Bull. Sch. Oriental Afr. Stud.* **64**, 240–247 (2001).
3. Chau, D. H. Woodblock printing, an essential medium of culture inheritance in chinese history. *J. Hong Kong Branch Royal Asiatic Soc.* 175–189 (1978).
4. Ma, H. The pre-emergence of chinese wood block printing—based on propagation medium and social transition. *Media Commun. Res.* **4**, 1–9 (2023).
5. Efros A A, F. W. T. Image quilting for texture synthesis and transfer. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 571–576 (2023).
6. Gatys L A, B. M., Ecker A S. Image quilting for texture synthesis and transfer. *Proc. IEEE Conf. on Comput. Vis. Pattern Recognition. Los Alamitos: IEEE Comput. Soc. Press.* 2414–2423 (2016).
7. Li S, N. L., Xu X. Laplacian-steered neural style transfer. *Proc. 25th ACM international conference on Multimed.* 1716–1724 (2017).
8. Risser, E., Wilmot, P. & Barnes, C. Stable and controllable neural texture synthesis and style transfer using histogram losses. *arXiv preprint arXiv:1701.08893* (2017).
9. Ulyanov, D., Lebedev, V., Vedaldi, A. & Lempitsky, V. Texture networks: Feed-forward synthesis of textures and stylized images. *arXiv preprint arXiv:1603.03417* (2016).
10. Gui X, Y. L. Y., Zhang B. Dlp-gan: learning to draw modern chinese landscape photos with generative adversarial network. *Neural computing & applications* **36**, 5267–5284 (2024).
11. Chen H, Z. L. Uatst: Towards unpaired arbitrary text-guided style transfer with cross-space modulation. *Comput. Vis. Image Underst.* **251**, 104246 (2025).
12. Xu, T., Yang, L. & Wu, E. Stroke-based real-time ink wash painting style rendering for geometric models. *SIGGRAPH Asia 2012 Tech. Briefs* (2012).
13. Yang, L. & Xu, T. Animating chinese ink painting through generating reproducible brush strokes. *Sci. China Inf. Sci.* **56**, 1–13 (2013).
14. He, M., Chen, D., Liao, J., Sander, P. V. & Yuan, L. Deep exemplar-based colorization. *ACM Transactions on Graph. (TOG)* **37**, 1 – 16 (2018).
15. Shrivastava, A. *et al.* Learning from simulated and unsupervised images through adversarial training. *2017 IEEE Conf. on Comput. Vis. Pattern Recognit. (CVPR)* 2242–2251 (2016).
16. Kang, X. *et al.* Ddcolor: Towards photo-realistic image colorization via dual decoders. *2023 IEEE/CVF Int. Conf. on Comput. Vis. (ICCV)* 328–338 (2022).
17. Afifi, M., Brubaker, M. A. & Brown, M. S. Histogan: Controlling colors of gan-generated and real images via color histograms. *2021 IEEE/CVF Conf. on Comput. Vis. Pattern Recognit. (CVPR)* 7937–7946 (2020).
18. Hong, K. *et al.* Aespa-net: Aesthetic pattern-aware style transfer networks. *2023 IEEE/CVF Int. Conf. on Comput. Vis. (ICCV)* 22701–22710 (2023).
19. Kolkin, N. I., Salavon, J. & Shakhnarovich, G. Style transfer by relaxed optimal transport and self-similarity. *2019 IEEE/CVF Conf. on Comput. Vis. Pattern Recognit. (CVPR)* 10043–10052 (2019).
20. Zhang, Y. *et al.* Domain enhanced arbitrary image style transfer via contrastive learning. *ACM SIGGRAPH 2022 Conf. Proc.* (2022).
21. Wu, X., Hu, Z., Sheng, L. & Xu, D. Styleformer: Real-time arbitrary style transfer via parametric style composition. *2021 IEEE/CVF Int. Conf. on Comput. Vis. (ICCV)* 14598–14607 (2021).
22. Deng, Y. *et al.* Stytr2: Image style transfer with transformers. *2022 IEEE/CVF Conf. on Comput. Vis. Pattern Recognit. (CVPR)* 11316–11326 (2021).
23. Park, D. Y. & Lee, K. H. Arbitrary style transfer with style-attentional networks. *2019 IEEE/CVF Conf. on Comput. Vis. Pattern Recognit. (CVPR)* 5873–5881 (2018).

24. Liu, S. *et al.* Adaattn: Revisit attention mechanism in arbitrary neural style transfer. *2021 IEEE/CVF Int. Conf. on Comput. Vis. (ICCV)* 6629–6638 (2021).

25. Huang, X. & Belongie, S. J. Arbitrary style transfer in real-time with adaptive instance normalization. *2017 IEEE Int. Conf. on Comput. Vis. (ICCV)* 1510–1519 (2017).

26. Huang, X., Liu, M.-Y., Belongie, S. J. & Kautz, J. Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision* (2018).

27. Jiang, L. *et al.* Tsit: A simple and versatile framework for image-to-image translation. *In ECCV* 2 (2020).

28. Jung, C., Kwon, G. & Ye, J. C. Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks. *In CVPR* 2,5,6 (2022).

29. Oza, M., Vaghela, H. & Bagul, S. Semi-supervised image-to-image translation. *2019 Int. Conf. Artif. Intell. Inf. Technol. (ICAIIT)* 16–20 (2019).

30. Jiang, Y., Jiang, L., Yang, S. & Loy, C. C. Scenimefy: Learning to craft anime scene via semi-supervised image-to-image translation. *2023 IEEE/CVF Int. Conf. on Comput. Vis. (ICCV)* 7323–7333 (2023).

31. Liu, Z., Luo, P., Wang, X. & Tang, X. Deep learning face attributes in the wild. *2015 IEEE Int. Conf. on Comput. Vis. (ICCV)* 3730–3738 (2014).

32. Chen, J., Liu, G. & Chen, X. Animegan: A novel lightweight gan for photo animation. In *International symposium on intelligence computation and applications*, 242–256 (Springer, 2019).

33. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A. & Torralba, A. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis machine intelligence* **40**, 1452–1464 (2017).

34. Zhang, R., Isola, P. & Efros. The unreasonable effectiveness of deep features as a perceptual metric. *In Proc. IEEE Conf. on Comput. Vis. Pattern Recognit. (CVPR)* 586–595 (2018).

35. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *In Adv. Neural Inf. Process. Syst. (NIPS)* 6626–6637 (2017).

36. Thyagarajan, K. S. & Gadre, V. Image processing: Principles and applications. *John Wiley Sons* (2013).

37. Wang, Z., Bovik, A., Sheikh, H. & Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Process.* 600–612 (2014).

38. Talebi, H. & Milanfar, P. Nima: Neural image assessment. *IEEE transactions on image processing* **27**, 3998–4011 (2018).

39. Kudo, M., Chung, H. & Osaki, Y. Prognostic staging system for hepatocellular carcinoma (clip score): its value and limitations, and a proposal for a new staging system, the japan integrated staging score (jis score). *J. gastroenterology* **38**, 207–215 (2003).

40. Wang, Z., Zhao, L. & Xing, W. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7677–7689 (2023).

41. Yang, L. *et al.* Cross-modal contextualized diffusion models for text-guided visual generation and editing. In *The Twelfth International Conference on Learning Representations* (2024).

42. Mahmood A, A. S., Bennamoun M. Resfeats: Residual network based features for image classification. In *2017 IEEE international conference on image processing (ICIP)*, 1597–1601 (2017).

43. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition (2014).

44. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).