Enhanced Image Restoration and Aesthetic Evaluation Using Modified YOLOv5 and Multi-Scale Residual Gated Convolutional Networks

Chang Jing
Henan College Of Transportation, Zhengzhou 451460, China
E-mail: m15939188369@163.com

Keywords: YOLOv5, MRGC, image restoration, image aesthetic evaluation, emotional fusion

Received: July 9, 2025

Cultural heritage represents humanity's invaluable treasure, yet existing image restoration methods suffer from inadequate edge feature extraction capabilities and aesthetic evaluation approaches that cannot be co-trained with emotional factors. Consequently, this research proposes an image restoration method based on an enhanced You Only Look Once version 5 small (YOLOv5s) and multi-scale residual fusion gated convolutions, integrating an aesthetic evaluation model that incorporates emotional elements. This approach employs the enhanced YOLOv5s for image extraction, utilizing the Canny operator as the edge detection algorithm. It incorporates multi-scale residual blocks and gated convolutional networks within the generative adversarial network, employing pixel reconstruction, perceptual, and style loss as the joint loss function. The research inserts emotion label extraction and emotion fusion modules into the residual network. Experiments demonstrate that the enhanced YOLOv5s achieves a maximum image extraction accuracy of 95.3%, surpassing both YOLOv5s and YOLOv8 by 12.5% and 0.3% respectively, whilst converging significantly faster. The image restoration model exhibits higher structural similarity indices, with restored images most closely approximating reality at a maximum value of 94.5%. Removing multi-scale residuals substantially impacts model performance. The aesthetic evaluation model achieves a maximum Spearman's correlation coefficient of 0.792 with the lowest computational complexity. Consequently, the proposed methodology effectively enhances image restoration capabilities and aesthetic evaluation quality, thereby facilitating the wider dissemination of cultural heritage.

Povzetek: Obnove slik kulturne dediščine so izvedene z izboljšanim YOLOv5 in večmerilno residualno vodenimi konvolucijami ter estetsko ocenjevanje z vključitvijo čustvenih dejavnikov. Pristop izboljša zaznavanje robov, vizualno kakovost in skladnost estetske ocene s človeško zaznavo.

1 Introduction

China stands as a venerable civilization boasting a rich and extensive history, during which it has amassed an incalculable wealth of cultural heritages over the ages. As time marches on, however, these invaluable treasures face inevitable degradation, with their external appearances and internal structures suffering damage from water stains, insect infestations, and the erosive forces of wind and rain [1]. Hence, it becomes imperative to undertake restoration efforts, for cultural heritage restoration transcends the mere act of repairing aged artifacts; it is a vital undertaking intricately linked to the perpetuation of human civilization, the affirmation of cultural identity, and the transmission of spiritual heritage [2]. Restoring cultural heritage can preserve material witness and evidence of history, prevent historical rifts, protect cultural diversity, continue cultural memory and identity, and maintain its artistic and aesthetic value. Within the domain of cultural heritage restoration, image restoration has always played a crucial role [3]. For paper cultural relics such as books and letters, image restoration can accurately predict missing parts through changes in the texture and color characteristics of the paper, thereby achieving the goal of restoration [4]. For mural cultural relics, image restoration algorithms can utilize the intact mural patterns and color information around damaged areas to generate images that closely resemble the original style and content. These algorithms offer valuable references for craftsmen during restoration, thereby helping to slow down the rate of mural deterioration. [5]. Meanwhile, aesthetic evaluation of cultural heritage images can select parts with important aesthetic value for key protection and inheritance, promoting people's understanding of their unique aesthetic characteristics and cultural connotations [6]. However, existing image restoration methods still suffer from insufficient ability to extract edge features, and aesthetic evaluation methods cannot train aesthetics and emotions together.

To deal with the problem of image restoration quality, Sun X et al. introduced a structure guided virtual restoration approach to address the lack of structural trends and poor performance of existing restoration algorithms. This approach incorporated an adaptive curve-fitting algorithm to rebuild missing structural lines. It devised a novel priority function to refine the sequence in which patches for repair were filled, and utilized an

adaptive approach for choosing sample block sizes according to structural sparsity [7]. The experiment showed that the average SSIM value of the repaired image using this approach was 0.977, and the average PSNR value was 39.16. Xu W et al. introduced a new color restoration technique grounded in the DenseNet algorithm to accurately restore the appearance of ancient relics and reduce the burden of manual restoration. This technology took a dataset consisting of 60 typical murals as system input and enhanced it through the DenseNet algorithm. Experiments showed that this technology was 44.62% faster than the SegNet algorithm in terms of time efficiency, 1.289% lower in structural similarity values than SegNet, 2.442% lower than Deeplab v3, and 1.288% lower than ResNet [8]. Xu H et al. introduced a new virtual reality fusion restoration approach for murals grounded in visual attention mechanism to address issues such as poor semantic correlation in mural restoration. This approach adopted the fusion technology of computation and perception to achieve the fusion association between virtual restored murals and the real spatial environment. It used a spatially layered and consistent detection approach to determine the fusion area between virtual and real. Experiments showed that this approach could effectively improve the efficiency of mural restoration and enhance the structural similarity of restored images [9]. Tang et al. proposed a novel multi-modal enhanced U-Net model to address the issue of substandard quality in real-world image restoration. This model leveraged pre-trained multi-modal large language models to extract semantic information from low-quality images, while employing an image encoder to enhance feature extraction capabilities. At the visual level, it achieved high-precision restoration through meticulous management of pixel-level spatial structures, integrating control information via a multi-layer attention mechanism. Experiments demonstrated that this model enabled precise and controllable image restoration [10]. Li L et al. introduced a new image aesthetic evaluation model grounded in theme perception visual attribute inference, which addressed the intrinsic relationship between visual attributes and image aesthetic quality in the evaluation of image aesthetic quality. This model simulated the human perception process of image aesthetics through two-layer inference, extracting aesthetic attribute features and thematic features separately, and introducing a flexible aesthetic network to extract general aesthetic features. Experiments showed that the introduced model outperformed existing state-of-the-art approaches and had better interpretability in four publicly available image aesthetic evaluation databases [11]. Celona L et al. introduced a new image aesthetic automatic prediction approach grounded in image analysis to address the issue of subjective influence on image aesthetic evaluation. This approach used a pre trained network to extract semantic features, used a multi-layer perceptron network to predict image attributes, and then generated pre encoded attribute information. Experiments showed that this approach could predict the style and composition attributes of different images, and calculate the distribution of aesthetic scores [12]. Yang Y et al. introduced a new conditional image aesthetic evaluation model to address the highly subjective issue in personalized image aesthetic evaluation. This model constructed a new personalized image aesthetics database, annotated by 438 subjects, and desensitized the image information. The subject information was used as a conditional prior to construct a conditional image aesthetics evaluation model. Experiments showed that this model could effectively address the evaluation limitations caused by annotation diversity, while maintaining the accuracy of image aesthetic evaluation [13]. In recent years, deep learning has achieved significant progress in the field of image restoration. Transformer architectures have demonstrated formidable potential in image restoration, such as the Swin Transformer effectively capturing long-range dependencies through its sliding window mechanism. Concurrently, diffusion models excel in generating high-quality images, offering novel approaches for cultural heritage restoration. Regarding aesthetic evaluation, visual Transformer-based approaches better model global aesthetic features through self-attention mechanisms. Nevertheless, these methods still exhibit limitations in adaptability to specific cultural heritage scenarios and the integration of emotional factors.

In summary, existing research has explored the issues of image restoration and aesthetic evaluation from multiple perspectives, and has achieved certain results. However, existing image restoration approaches still have problems such as repairing image variation and blur, and aesthetic evaluation approaches have a single emotional element. Therefore, this study proposes an image restoration and aesthetic evaluation method based on Multi-scale Residuals and Gated Convolution (MRGC) and Aesthetic Evaluation of Emotional Integration (AEEI). The research aims to enhance the detection accuracy of cultural heritage images in complex backgrounds and improve the precision of image modifications. Additionally, it seeks to overcome the edge blurring issues inherent in traditional methods and integrate subjective factors, such as emotional evaluation, into aesthetic assessment models. This approach brings evaluation outcomes closer to human subjective perception, elevates the quality of image aesthetic evaluation, and strengthens public cultural identity. The method innovatively proposes an improved YOLOv5s architecture, optimized through the C3-DSC module and attention mechanisms, significantly enhancing the detection accuracy and efficiency of cultural heritage targets. A multi-scale residual fusion gated convolutional network is designed to effectively resolve edge blurring in cultural heritage image restoration. For the first time, an emotion label fusion mechanism is introduced into the aesthetic evaluation of cultural heritage images, achieving collaborative modelling of aesthetic attributes and emotional characteristics. The superiority of the proposed method is validated across multiple cultural heritage datasets, providing a novel technical pathway for intelligent cultural heritage preservation.

2 Methods and materials

2.1 Target image extraction based on improved YOLOv5

When extracting cultural heritage images, it is necessary to identify the cultural heritage in the image clearly due to the influence of lighting intensity and occlusion, as well as the texture details of the target object [14]. The research adopts the YOLOv5s algorithm as the basic structure, but the ordinary YOLOv5s algorithm has problems such as difficulty in recognizing complex targets and high model complexity. Therefore, research has been conducted to improve it, and the distinctive configuration of the improved YOLOv5s model is presented in Figure 1.

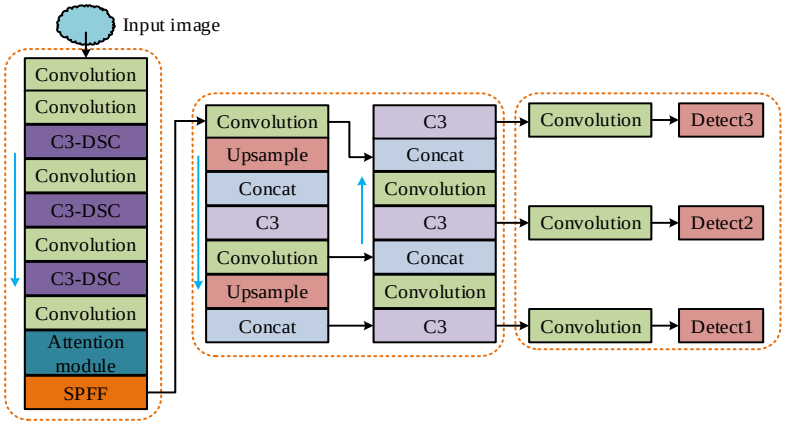

Figure 1: Improved distinctive configuration of the YOLOv5s model (Image source: Authors own illustration)

In Figure 1, the improved YOLOv5s model mainly consists of three parts: the backbone network (BN), the neck network, and the detection end. The BN includes a convolution module, a C3-Deepwise Separable Convolution (C3-DSC) module formed by fusing C3 and depthwise separable convolutions, an attention mechanism module, and spatial pyramid fast pooling. The C3-DSC module can effectively reduce the number of parameters in the model, improve computing speed, and the attention module can model in the channel dimension, highlighting important feature channels and suppressing unimportant channels, thereby enhancing the model's attention to key features. In addition to the above modules, the neck network also adds an Upsample module to upsample the feature maps (FMs), and a Concat module to concatenate feature images from different levels. After passing through the convolution module, the detection data are input into the Detect module to predict bounding boxes on FMs of different scales.

The C3 module has two parallel paths, with the main path consisting of one-dimensional convolution, multiple Bottleneck residual blocks, and one-dimensional convolution. The shortcut path only has one-dimensional convolution, and the outputs of the two paths are concatenated in the channel dimension. The Bottleneck residual block is composed of two successive 3D convolutions. In this study, the optimization of the Bottleneck block is achieved solely by substituting its 3D convolution with a depthwise separable convolution. The attention module enables the model to focus its attention on the critical features of the target. Moreover, the study incorporated just a single layer into the BN, which serves to enhance the image's feature extraction capabilities. The attention module can achieve cross channel transmission of data with fewer parameters. To prevent dimensionality reduction, one-dimensional convolution is introduced in the attention module. The adaptive function calculation used is presented in equation (1) [15].

$$A=\varphi(C)=\left|\frac{\log_2^C}{\lambda}+\frac{b}{\lambda}\right|_{odd} \quad (1)$$

In equation (1), A represents the output result, $\varphi(C)$ represents the adaptive function, $|\cdot|_{odd}$ represents rounding the solution down to the nearest odd number, λ represents the constant for adjusting the scaling ratio, b represents the constant for adjusting the calculation offset, and C represents the channel dimension. The distinctive configuration of the attention module is presented in Figure 2.

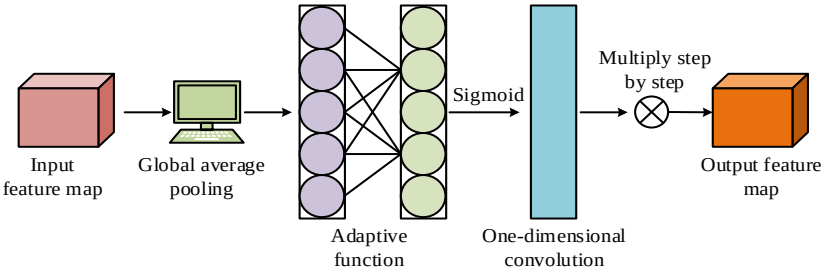

Figure 2: The distinctive configuration of the attention module (Image source: Authors own illustration)

In Figure 2, the FM is first inputted and globally average pooled to aggregate spatial information. Then, an adaptive function is used to determine channel attention weights. After the attention module, one-dimensional convolution is employed to generate attention weights for each channel. The Sigmoid activation function is used for linear transformation, and finally the generated attention weights are multiplied step by step with the original FM to obtain a weighted FM. The dimensions of the output FM stay consistent. To reduce the total degree of freedom of the LF, the SIoU LF is used to replace the traditional CIoU LF. The introduction of angle loss term in SIoU can guide the model to correct directional deviation and introduce distance decoupling penalty, making it more sensitive to small target offset. Adaptive shape penalty performs position correction in the initial stage and shape fine-tuning in the later stage. IoU loss provides clear gradient direction and reduces oscillation.

2.2 Image restoration based on MRGC model

The improved YOLOv5s algorithm used in the study can only extract target images, but most existing cultural heritage objects suffer from problems such as missing patterns, textures, or colors [16]. Therefore, research needs to repair the extracted images and reconstruct the missing areas of the images. However, existing end-to-end deep learning image restoration methods still suffer from the problem of restoring blurry image edges. Therefore, the MGRC model that combines multi-scale residuals (MSRs) with gated convolutional networks is studied. The distinctive configuration of the model is presented in Figure 3.

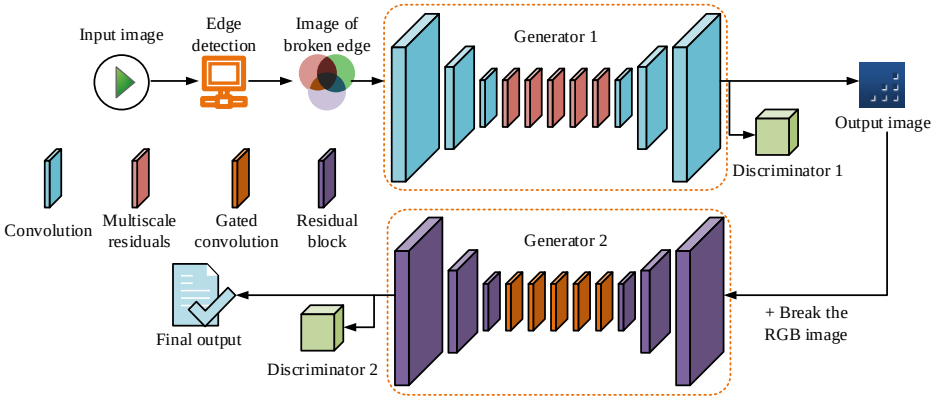

Figure 3: Distinctive configuration of the MGRC model (Image source: Authors own illustration)

In Figure 3, the model mainly consists of three parts: edge detection module, edge generation module, and texture restoration module. The main structure of the second and third parts is similar to that of a generative adversarial network, both containing a generator and a discriminator. The MGRC model inputs the extracted images into the edge generation module, which constructs the initial image through the generator. Then, it is jointly trained with the discriminator to generate images with higher realism and input them into the texture restoration module. The input of the texture restoration module also includes the Red Green Blue (RGB) image of the initial image, which is iteratively trained using a generator and discriminator. The generator of the edge generation module consists of multiple convolutions and MSR modules, while the generator of the texture restoration module consists of multiple gated convolutions and residual modules. The study adopts the Canny operator as the edge detection algorithm, which can reduce noise interference and improve edge localization accuracy. The input images of the edge generation module include grayscale images and edge images. The image expression after calculation at the input end is presented in equation (2) [17].

$$\begin{cases} I_G = I_g \square \left(1-M\right) \\ I_E = I_e \square \left(1-M\right) \end{cases} \tag{2}$$

In equation (2), I_G represents the damaged grayscale image, I_g represents the grayscale image, $\square$ represents the Hadamard product, M represents the mask, I_E represents the broken edge image, and I_e represents the edge image. After the image calculation is completed, the study concatenates I_G, I_E, and the mask, and inputs them into the generator after concatenation. The generator extracts image features, performs broken area repair, and inputs the repaired edge image into the discriminator for iteration. The final generated repaired image is calculated as shown in equation (3).

$$I'_R = I_R \square M + I_E \tag{3}$$

In equation (3), I'_R represents the final repaired image and I_R represents the initial repaired image output by the generator. The study suggests that an MSR module can improve the completeness of feature extraction in the model and enhance the edge continuity of the repaired image. Its distinctive configuration is presented in Figure 4.

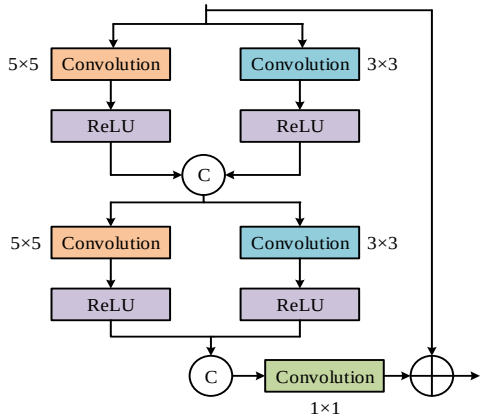

Figure 4: Distinctive configuration of the MSR module
(Image source: Authors own illustration)

In Figure 4, the MSR is a parallel dual branch structure, with two branches having the same structure. After three-dimensional convolution, small-scale local features of the image are gathered, and a five dimensional convolution is used to capture contextual information with a larger receptive field. ReLU activation function is reused to introduce non-linear processing capability. Two branch FMs are concatenated along the channel dimension to aggregate local features. Finally, the concatenated FMs are input into one-dimensional convolution to explore the correlation and complementary information of features at different scales. The input image of the generator consists of a damaged RGB image and I'_R stitching, calculated as shown in equation (4).

$$\begin{cases} I_{TG} = \left\{ I'_R, I'_{RGB} \right\} \\ I'_{RGB} = I_{RGB} \square \left(1 - M \right) \end{cases} \tag{4}$$

In equation (4), I_{TG} represents the input image of the generator, I'_{RGB} represents the broken RGB image, and I_{RGB} represents the initial RGB image. After multiple iterations in the generator and discriminator, the final output graph is calculated as shown in equation (5).

$$I'_{TR} = I_{TR} \square M + I'_{RGB} \tag{5}$$

In equation (5), I'_{TR} represents the final output image and I_{TR} represents the initial output image of the texture restoration module generator. To improve the authenticity of restored images, generative adversarial loss and feature matching loss were used as the LFs of the edge generation module. The calculation of generative adversarial loss is presented in equation (6) [18].

$$L_{ga} = \mathrm{E}_{(I_e, I_g)} \left[\log D_1 \left(I_e, I_g \right) \right] + \mathrm{E}_{I_g} \cdot \log \left[1 - D_1 \left(I'_R, I_g \right) \right] \tag{6}$$

In equation (6), L_{ga} is the generative adversarial loss, E is the mathematical expectation, and D_1 is the discriminator of the edge generation module. The calculation of feature matching loss is presented in equation (7).

$$L_{fm} = \mathrm{E} \left[\sum_{i=1}^{k} \frac{1}{n_i} \left\| d_i \cdot I_e - d_i \cdot I'_R \right\|_1 \right] \tag{7}$$

In equation (7), L_{fm} represents the feature matching loss, k represents the number of convolutional layers in the edge generation module discriminator, n_i represents the total number of elements in the discriminator activation layer, and i represents the layer activation map. The joint LF calculation of the edge generation module is presented in equation (8).

$$\min_{G_1} \max_{D_1} L_{G_1} = \min_{G_1} \left(\omega_{ga} \max_{D_1} \left(L_{ga} \right) + \omega_{fm} L_{fm} \right) \tag{8}$$

In equation (8), L_{G_1} represents the joint LF of the edge generation module, G_1 represents the generator of the edge generation module, ω_{ga} represents the weight coefficients for generating adversarial loss, and ω_{fm} represents the weight coefficients for feature matching loss. The LF of the texture restoration module includes four types: generative adversarial loss, pixel reconstruction loss, perceptual loss, and style loss. The calculation method of generative adversarial loss is the same as that of the edge generation module. The calculation of pixel reconstruction loss is presented in equation (9).

$$L_{pr} = \frac{1}{M} \left\| I'_{TR} - I_{RGB} \right\|_1 \tag{9}$$

In equation (9), L_{pr} represents the pixel reconstruction loss. The calculation of perceptual loss is presented in equation (10).

$$L_p = \mathrm{E}\left[\sum_{i=1}\frac{1}{n_i}\left\|Y_i \cdot I_{RGB} - Y_i \cdot I_{TR}\right\|_1\right] \quad (10)$$

In equation (10), L_p represents perceptual loss and Y_i represents the activation map of the pre trained network at the layer. The calculation of style loss is presented in equation (11).

$$L_s = \mathrm{E}\left[\left\|G_Y\left(I_{TR}\right) - G_Y\left(I_{RGB}\right)\right\|_1\right] \quad (11)$$

In equation (11), L_s represents style loss and G_Y represents the Gram matrix composed of activation graph Y_i. The joint LF calculation of the texture repair module is presented in equation (12).

$$\min_{G_2}\max_{D_2} L_{G_2} = \min_{G_2}\left(\omega_{ga'}\max_{D_2}\left(L_{ga'}\right) + \omega_{pr}L_{pr} + \omega_p L_p + \omega_s L_s\right) \quad (12)$$

In equation (12), G_2 represents the generator of the texture repair module, D_2 represents the discriminator of the texture repair module, L_{G_2} represents the joint LF of the texture repair module, $L_{ga'}$ represents the generative adversarial loss of the texture repair, $\omega_{ga'}$, ω_{pr}, ω_p, and ω_s represent the weight coefficients of generative adversarial, pixel reconstruction, perceptual, and style loss, respectively.

2.3 Aesthetic evaluation of restored images based on emotional fusion

Cultural heritage images often carry past aesthetic concepts and cultural information, and aesthetic evaluation of restored images can restore narrative vitality to vanished fragments of civilization [19]. Cultural heritage images often contain various emotional attributes, including positive and negative, but existing aesthetic evaluation methods for images cannot train aesthetics and emotions together. Therefore, the study developed an image aesthetic evaluation model grounded in emotional fusion, and the distinctive configuration of the model is presented in Figure 5.

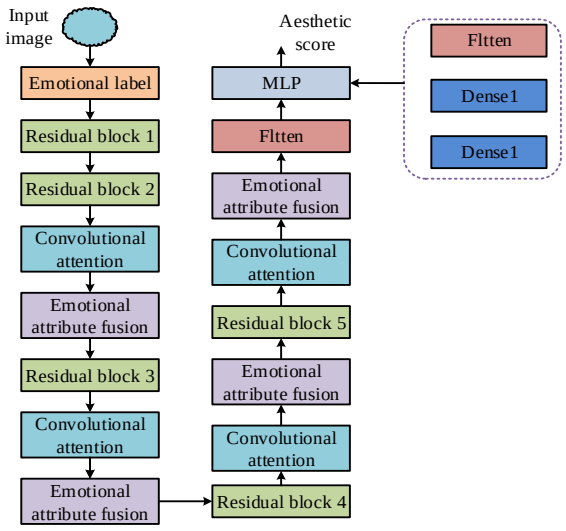

Figure 5: Image aesthetic evaluation model of emotional fusion
(Image source: Authors own illustration)

In Figure 5, the overall model is a linear structure with a ResNet50 network as the backbone structure, in which multiple modules are inserted. After image input, emotional labels are used for annotation, and feature extraction is performed after annotation is completed. The convolutional attention module composed of channel and spatial attention mechanisms is utilized to enhance feature expression ability. The study constructs an emotional attribute fusion module to integrate the extracted emotional and attribute labels into the model. After multiple rounds of feature extraction, the study uses the Fltten layer to convert multidimensional data into one-dimensional output data, and predicts aesthetic scores in a multi-layer perceptron. In aesthetic evaluation, the emotions contained in an image can affect its final rating. To introduce image emotions into the evaluation model, research is being conducted on preprocessing the input image using emotional labels. The label processing of images requires first obtaining label vectors of different emotions, performing an expanding operation on the label vectors, that is, expanding the low dimensional semantic vectors into high-dimensional spatial features, as shown in equation (13) [20].

$$F_{emo} = \Gamma(W \cdot emo + b) \quad (13)$$

In equation (13), F_{emo} is the output FM, Γ is the non-linear activation function, W is the weight matrix, emo is the sentiment label vector, and b is the bias term. After processing, the research will concatenate the output FM with the initial input image, and calculate as shown in equation (14).

$$F_i = F_{emo} \oplus F_{ima} \quad (14)$$

In equation (14), F_i is the input image annotated with emotional labels, $\oplus$ is feature concatenation, and F_{ima} is the initial input image. In the convolutional

attention module, the FMs are sequentially subjected to channel and spatial attention calculations, as shown in equation (15) [21].

$$\begin{cases} M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ M_s(F) = \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)])) \end{cases} \tag{15}$$

In equation (15), $M_c(F)$ is the output of channel attention, σ is the Sigmoid activation function, MLP is the multi-layer perceptron, $AvgPool$ is global average pooling, F is input features, and $MaxPool$ is global max pooling. $M_s(F)$ is the output of the spatial attention, and $f^{7\times7}$ is a 7×7 convolution. The distinctive configuration of the emotional attribute fusion module is presented in Figure 6.

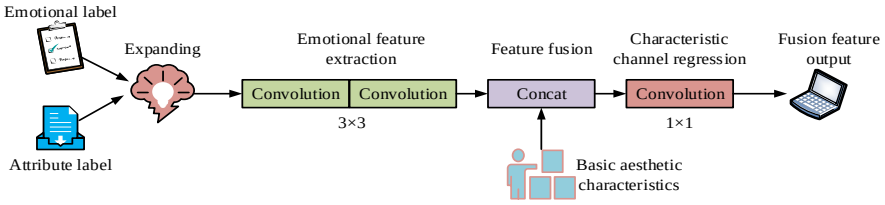

Figure 6: Distinctive configuration of the emotional attribute fusion module
(Image source: Authors own illustration)

In Figure 6, after being input into the fusion module, the image undergoes the addition of emotional and attribute labels, followed by the expansion operation. The attribute labels are the degree of color vividness, color harmony, and element balance, respectively. After processing, the image fusion features are obtained and subjected to two 3D convolutions to extract emotional features, which are then fused with basic aesthetic features. Finally, one-dimensional convolution was used for feature regression processing and output to the residual module of ResNet50 network.

3 Results

3.1 Analysis of image restoration experiment

In the hardware environment of the experiment, the CPU was AMD Ryzen 7, the GPU was NVIDIA RTX 3090, the video memory was 24GB, the memory was 64GB DDR4, and the storage was 1TB SSD. The operating system in the software environment was Ubuntu 22.04 LTS, and the deep learning framework was PyTorch 1.12.0. Among the model's parameter settings, the batch size was set to 32, the initial learning rate to 0.001, the cosine annealing scheduler employed, and the minimum learning rate to 1e-6. The optimizer utilized AdamW with a weight decay of 0.01. Training spanned 100 epochs, with early stopping triggered when the validation set loss failed to decrease for 10 consecutive epochs. The experimental datasets comprised PSI-Art and Dunhuang Grottos. The PSI-Art dataset contained 12,500 images of European murals and sculptures, encompassing varying degrees of damage. The Dunhuang Grottos dataset comprised 8,760 images of murals from the Mogao Caves in Dunhuang, featuring rich colour and texture characteristics. To ensure fair evaluation, the datasets were randomly partitioned into training, validation, and test sets at an 8:1:1 ratio. Data augmentation applied random horizontal flipping (probability 0.5), random rotation (±20°), and colour dithering with brightness, contrast, and saturation adjusted by 0.2 each. The comparison of different LF curves for improving YOLOv5s is presented in Figure 7.

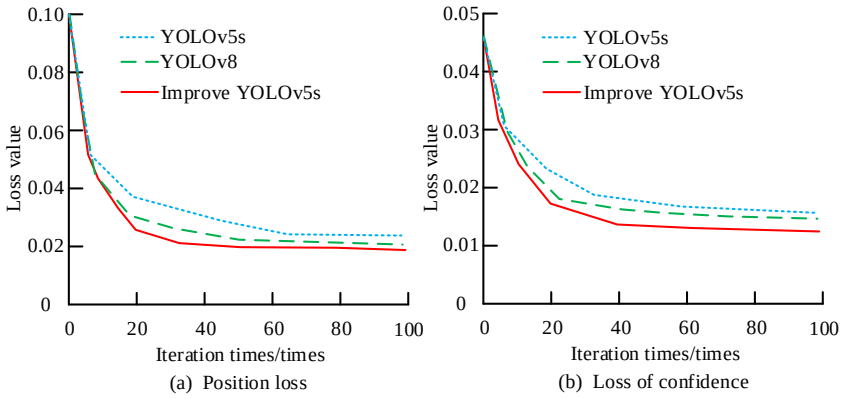

Figure 7: Comparison of different LF curves of the improved YOLOv5s
(Image source: Authors own illustration)

In Figure 7 (a), the three different algorithms had a small difference in iteration speed in the early stage, but the improved YOLOv5s algorithm converged faster after 10 iterations. The minimum position loss value of the improved YOLOv5s algorithm was 0.019, which was 0.009 and 0.004 lower than YOLOv5s and YOLOv8, respectively. In Figure 7(b), owing to the integration of the attention module into the model, the enhanced YOLOv5s exhibited a quicker convergence rate for confidence loss, achieving a minimum confidence loss value of 0.014. This value was 0.005 and 0.003 lower than those of YOLOv5s and YOLOv8, respectively. The comparison of the accuracy of target image extraction using the improved YOLOv5s algorithm in different datasets is presented in Figure 8.

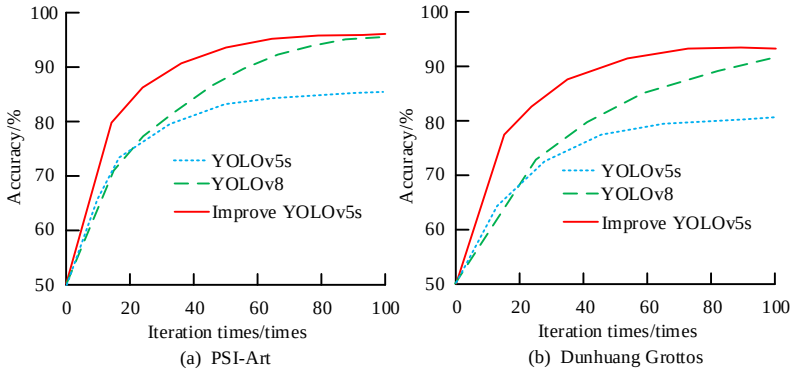

Figure 8: Accuracy rate of target image extraction of the improved YOLOv5s algorithm in different datasets (Image source: Authors own illustration)

In Figure 8 (a), the image extraction accuracy of improved YOLOv5s was the highest, with a maximum value of 95.3%, which was 12.5% and 0.3% higher than YOLOv5s and YOLOv8, respectively, and the convergence speed was significantly better than the other two algorithms. In Figure 8 (b), the images in the Mogao Grottoes mural dataset were more complex and fragmented. The extraction accuracy of all three algorithms decreased, with the improved YOLOv5s algorithm decreasing by 2.8%, YOLOv5s and YOLOv8 decreasing by 3.5% and 3.2%. The comparison of image restoration performance among various models is presented in Table 1.

Table 1: Comparison of image restoration performance of various models

Indicator	SSIM/%			L1/%		
Mask ratio	[0-0.2)	[0.2-0.4)	[0.4-0.6)	[0-0.2)	[0.2-0.4)	[0.4-0.6)
MRGC	94.5	85.1	70.3	1.92	3.84	7.52
CA	87.3	72.5	59.6	3.85	6.92	11.08
CTSDG	90.8	79.3	65.7	1.99	5.02	8.15
GANs	85.4	70.2	55.4	4.19	7.25	12.44
BM3D	89.2	75.5	63.8	2.28	5.52	9.17
MAT	91.7	82.4	68.2	1.97	4.28	8.10
IPT	92.5	82.9	68.5	1.95	4.13	8.04

In Table 1, the comparison algorithms are Contextual Attention (CA), Condition Texture and Structure Dual Generation (CTSDG), Generative Adversarial Networks (GANs), Block Matching and 3D Filtering (BM3D), Mask-Aware Transformer (MAT), and Image Processing Transformer (IPT). The MRGC model achieved the highest structural similarity index, producing restored images closer to reality. When the mask ratio fell within the range [0–0.2), the model attained an SSIM value of 94.5%, surpassing the second-best model, IPT, by 2.0%. Its L1 value stood at 1.92%, which was 0.03% and 0.05% lower than those of the IPT and MAT models respectively. The comparison of MRGC model ablation experimental results is presented in Figure 9.

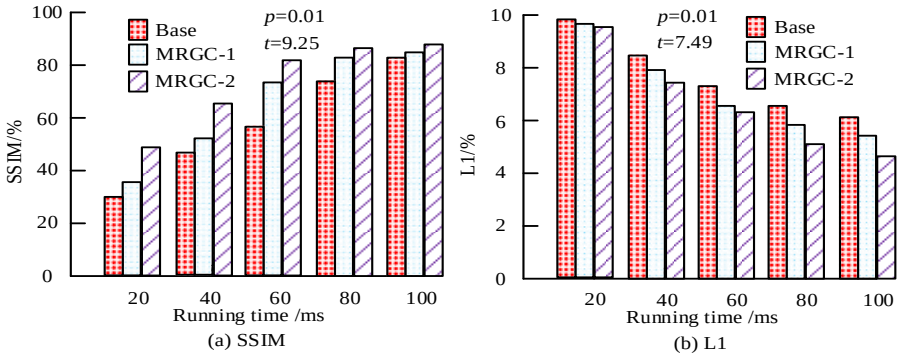

Figure 9: Comparison of ablation experimental results of the MRGC model
(Image source: Authors own illustration)

In Figure 9 (a), to validate the statistical significance of the results, five independent experiments were conducted for each module of the model ablation. The experimental outcomes represented average performance metrics, with t-tests indicating statistically significant differences between experimental results ($p < 0.05$). Base represents the removal of MSRs and gated convolutions, MRGC-1 represents the removal of MSRs, and MRGC-2 represents the removal of gated convolutions. The performance of MRGC-2 was significantly better than MRGC-1 and Base, with a maximum SSIM value of 88.2%, which was 4.8% and 6.9% higher than MRGC-1 and Base, respectively. In Figure 9 (b), the minimum L1 norms of Base, MRGC-1, and MRGC-2 were 6.25%, 5.87%, and 4.92%, respectively.

3.2 Experimental analysis of aesthetic evaluation of restored images

The basic settings of the aesthetic evaluation model were the same as in section 2.1, with an initial learning rate of 0.0001 and a max iteration count of 20. The comparative models used in the experiment included Hierarchical Layout Aware Graph Convolutional Network (HLA-GCN), Self-Supervised Vision Transformer (SSViT), Adaptive Fractional Dilation Convolution (AFDC), and Aesthetic Attribute Prediction Network (AttributeNet). The performance comparison of various models is presented in Figure 10.

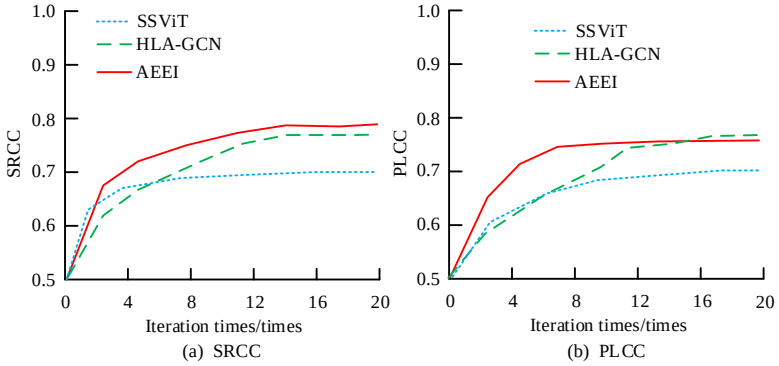

Figure 10: Performance comparison of different models
(Image source: Authors own illustration)

In Figure 10 (a), SRCC is the Spearman rank correlation coefficient, and PLCC is the Pearson linear correlation coefficient. The larger the coefficient, the closer the model's predicted value is to the true value. The maximum SRCC coefficient of the AEEI model proposed by the research was 0.792, which was 0.021 and 0.108 higher than HLA-GCN and SSViT, respectively. In Figure 10 (b), the PLCC coefficient of the AEEI model tended to converge after 8 iterations, with a maximum coefficient value of 0.760, slightly lower than HLA-GCN. The comparison of computational complexity among various models is presented in Table 2.

Table 2: Computational complexity of different models

Model	Parameter quantity	GFLOPs	Memory usage (KB)	Runtime (ms)	Time complexity(O^2)
AEEI	14268	127.5	263	67	Lower
HLA-GCN	354862	56.8	720	94	Normal
SSViT	59894	169.4	1806	285	Higher

AFDC	105249	25.7	5307	1052	Higher
Attribute Net	82183	11.2	2125	128	Normal

In Table 2, the AEEI model had the lowest computational complexity with 14268 parameters, 127.5 floating-point calculations, and 263KB of memory usage, which was 457KB lower than the second best HLA-GCN model. The running time of the AEEI model was 67ms, which was better than other models and had lower time complexity.

4 Conclusion

Given the limitations in edge feature extraction capabilities and the inability of existing image restoration methods to jointly train aesthetics and emotions in their aesthetic evaluation approaches, this study introduces an image restoration and aesthetic evaluation method that leverages MRGC (Multi-Resolution Graph Convolution) and emotion fusion. The experiment showed that the minimum position loss value and confidence loss value of the improved YOLOv5s algorithm were 0.019 and 0.014, respectively, which were lower than other algorithms. The improved YOLOv5s had the highest image extraction accuracy, with a maximum value of 95.3%, which was 12.5% and 0.3% higher than YOLOv5s and YOLOv8, respectively, and the convergence speed was significantly better than the other two algorithms. The MRGC model had a higher structural similarity index in the repaired image, and the repaired image was closest to the real situation. When the mask ratio was between [0-0.2), the SSIM value of the MRGC model was 94.5%, which was 3.7% higher than the second best CTSDG. Removing MSRs had a greater impact on the performance of the MRGC model, with the maximum SSIM value of MRGC-2 being 88.2%, which was 4.8% and 6.9% higher than MRGC-1 and Base, respectively. The maximum SRCC coefficient of the AEEI model was 0.792, which was 0.021 and 0.108 higher than HLA-GCN and SSViT. The PLCC coefficient tended to converge after 8 iterations, with a maximum coefficient value of 0.760, slightly lower than HLA-GCN. The AEEI model had the lowest computational complexity, with a parameter count of 14268, floating-point calculations of 127.5, and memory usage of 263KB, which was 457KB lower than the second best HLA-GCN model. The MRGC model's outstanding performance in structural similarity demonstrated its ability to effectively preserve the original structural and textural characteristics of cultural heritage images, which was crucial for safeguarding the historical authenticity of artefacts. Compared to current state-of-the-art Transformer-based restoration methods, this approach achieved a superior balance between computational complexity and restoration quality, making it particularly well-suited for resource-constrained cultural heritage conservation scenarios. This study also presents certain limitations, such as the restoration quality of the MRGC model being partially dependent on the accuracy of the front-end edge generation module, and the annotation of sentiment labels inherently possessing a degree of subjectivity. Future research could explore more robust edge detection and generation algorithms, investigate the application of multi-modal information (such as textual descriptions) in aesthetic evaluation, and extend the proposed methodology to the restoration and assessment of three-dimensional cultural heritage models.

Funding statement

This study was supported by A Study on the Inheritance and Innovation of Traditional Chinese Aesthetics, Soft Science Research Project, Henan Provincial Department of Science and Technology, Project Approval Number: 252400411131.

List of variables:

A: Output result

$\varphi(C)$: Adaptive function

$|\cdot|_{odd}$: Rounding the solution down to the nearest odd number

λ: Constant for adjusting the scaling ratio

b: Constant for adjusting the calculation offset

C: Channel dimension

I_G: Damaged grayscale image

I_g: Grayscale image

$\odot$: Hadamard product

M: Mask

I_E: Broken edge image

I_e: Edge image

I'_R: Final repaired image

I_R: Initial repaired image output by the generator

I_{TG}: Input image of the generator

I'_{RGB}: Broken RGB image

I_{RGB}: Initial RGB image

I'_{TR}: Final output image

I_{TR}: Initial output image of the texture restoration module generator

L_{ga}: Generative adversarial loss

E: Mathematical expectation

D_1: Discriminator of the edge generation module

L_{fm}: Feature matching loss

k: Number of convolutional layers in the edge generation module discriminator

n_i: Total number of elements in the discriminator activation layer

i: Layer activation map

L_{G_1}: Joint LF of the edge generation module

G_1 : Generator of the edge generation module

ω_{ga} : Weight coefficients for generating adversarial loss

ω_{fm} : Weight coefficients for feature matching loss

L_{pr} : Pixel reconstruction loss

L_p : Perceptual loss

Y_i : Activation map of the pre trained network at the layer

L_s : Style loss

G_Y : Gram matrix composed of activation graph Y_i

G_2 : Generator of the texture repair module

D_2 : Discriminator of the texture repair module

L_{G_2} : Joint LF of the texture repair module

F_{emo} : Output FM

Γ : Non-linear activation function

W : Weight matrix

emo : Sentiment label vector

b : Bias term

F_i : Input image annotated with emotional labels

$\oplus$: Feature concatenation

F_{ima} : Initial input image

$M_c(F)$: Output of channel attention

σ : Sigmoid activation function

MLP : Multi-layer perceptron

$AvgPool$: Global average pooling

F : Input features

$MaxPool$: Global max pooling

$M_s(F)$: Output of the spatial attention

$f^{7\times7}$: 7×7 convolution

Funding

This study was supported by A Study on the Inheritance and Innovation of Traditional Chinese Aesthetics, Soft Science Research Project, Henan Province Soft Science Research Program Project, Project Approval Number: 252400411131.

References

[1] Pu H, Wang X. The impact of environment on cultural relics. Scientific Culture, 2023, 9(2): 12-25. DOI: 10.2307/j. Ctv141632p.

[2] Gao Z, Du M, Cao N, Hou M, Wang W, Lyu S. Application of hyperspectral imaging technology to digitally protect murals in the Qutan temple. Heritage Science, 2023, 11(1): 8-27. DOI: 10.1186/s40494-022-00847-7.

[3] Chang, H., & Ding, Q. (2025). Hierarchical local-global attention in a multi-scale transformer network for enhanced image denoising. Informatica, 49(6): 9-15. DOI: 10.15388/22-INFOR480.

[4] Vera Nieto D, Celona L, Fernandez Labrador C. Understanding aesthetics with language: A photo critique dataset for aesthetic assessment. Advances in Neural Information Processing Systems, 2022, 35(7): 34148-34161. DOI: 10.1145/3539618.3591817.

[5] Satvati, M. A., Lakestani, M., Khamnei, H. J., & Allahviranloo, T. (2024). Deblurring Medical Images Using a New Grünwald-Letnikov Fractional Mask. Informatica, 35(4), 817-836. DOI: 10.15388/24-infor573.

[6] Wang Y, Song F, Liu Y, Li Y, Ma X, Wang W. Research on the correlation mechanism between eye - tracking data and aesthetic ratings in product aesthetic evaluation. Journal of engineering design, 2023, 34(1): 55-80. DOI: 10.1080/09544828.2023.2172662.

[7] Sun X, Jia J, Xu P, Ni J, Shi W, Li B. Structure - guided virtual restoration for defective silk cultural relics. Journal of Cultural Heritage, 2023, 62(2): 78-89. DOI: 10.1016/j. Culher.2023.05.016.

[8] Xu W, Fu Y. Deep learning algorithm in ancient relics image colour restoration technology. Multimedia Tools and Applications, 2023, 82(15): 23119-23150. DOI: 10.1007/s11042 - 022 - 14108 - z.

[9] Xu H, Zhang Y, Zhang J. Frescoes restoration via virtual - real fusion: Method and practice. Journal of Cultural Heritage, 2024, 66(15): 68-75. DOI: 10.1016/j. Culher.2023.11.001.

[10] Tang A, Wei L, Ni Z, & Huang Q. (2025). Multi-Modal Modified U-Net for Text-Image Restoration: A Diffusion-Based Multimodal Information Fusion Approach. Informatica, 49(2): 12-23. DOI: 10.31449/inf.v49i2.8245.

[11] Li L, Huang Y, Wu J, Yang Y, Li Y, Guo Y, Shi G. Theme - aware visual attribute reasoning for image aesthetics assessment. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(9): 4798-4811. DOI: 10.1109/tcsvt.2023.3249185.

[12] Celona L, Leonardi M, Napoletano P, Rozza A. Composition and style attributes guided image aesthetic assessment. IEEE Transactions on Image Processing, 2022, 31(17): 5009-5024. DOI: 10.1109/tip.2022.3191853.

[13] Yang Y, Xu L, Li L, Qie N, Li Y, Zhang P, Guo Y. Personalized image aesthetics assessment with rich attributes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 7(2): 19861-19869. DOI: 10.1109/cvpr52688.2022.00305.

[14] Jung H K, Choi G S. Improved yolov5: Efficient object detection using drone images under various conditions. Applied Sciences, 2022, 12(14): 7255-7279. DOI: 10.25236/ajcis.2023.061202.

[15] Papenmeier F, Dagit G, Wagner C, Schwan S. Is it art? Effects of framing images as art versus non - art on gaze behavior and aesthetic judgments.

Psychology of Aesthetics, Creativity, and the Arts, 2024, 18(4): 642-657. DOI: 10.1037/aca0000466.

[16] Hasanvand M, Nooshyar M, Moharamkhani E, and Selyari A. Machine learning methodology for identifying vehicles using image processing. AIA, 2023, 1(3): 170-178. DOI: 10.47852/bonviewaia3202833.

[17] Zhang Z, Sun W, Zhou Y, Jia J, Zhang Z, Liu J, et al. Subjective and objective quality assessment for in - the - wild computer graphics images. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 20(4): 1-22. DOI: 10.1145/3631357.

[18] Qiu S, Zhang P, Tang X, Zeng Z, Zhang M, Hu B. Sanxingdui cultural relics recognition algorithm based on hyperspectral multi - network fusion. Computers, Materials & Continua, 2023, 77(3): 19-35. DOI: 10.32604/cmc.2023.042074.

[19] Sha S, Li Y, Wei W, Liu Y, Chi C, Jiang X, et al. Image classification and restoration of ancient textiles based on convolutional neural network. International Journal of Computational Intelligence Systems, 2024, 17(1): 11-27. DOI: 10.31219/osf. Io/kbrjf.

[20] Tao N, Wang C, Zhang C, Sun J. Quantitative measurement of cast metal relics by pulsed thermal imaging. Quantitative InfraRed Thermography Journal, 2022, 19(1): 27-40. DOI: 10.21611/qirt.2019.019.

[21] Hu C, Huang X, Xia G, Liu X, Ma X. A high - precision automatic extraction method for shedding diseases of painted cultural relics based on three - dimensional fine color model. Heritage Science, 2024, 12(1): 300-325. DOI: 10.21203/rs.3. Rs-3804835/v1.