A Hybrid CNN-GrabCut and GAN Architecture for Semantic Segmentation and Artistic Style Transfer in Image Generation

Xin Zhou
Shangqiu Normal University, Shangqiu 476000, China
E-mail: 15082955511@163.com

Keywords: Convolutional neural network, GrabCut, style transfer, semantic segmentation

Received: June 13, 2025

Driven by digital transformation and the growing demand for artistic innovation, the development of artistic image generation systems has gained wide attention. However, existing systems face several limitations, including low accuracy in semantic segmentation, unnatural fusion in style transfer, and weak originality in generated images. Therefore, an artistic image generation system model based on graph cutting method convolutional neural network generative adversarial network is proposed, which uses an improved convolutional neural network (CNN) combined with GrabCut algorithm for semantic segmentation. By introducing convolutional block attention module to optimize CNN, it can better capture global information and complex features. In terms of style transfer, the system uses a semantic guided GAN architecture to achieve precise adaptation and natural transition of style features. The study achieved a pixel accuracy of 98% on the PASCAL VOC dataset. With the increase of training data, its highest intersection to union ratio reached 92%, significantly higher than the comparison algorithm. In real-time testing, the system only occupies 710 MB of memory and has a response time of less than 62 ms, outperforming other models in performance. In addition, in the image quality test conducted on the ArtBench-10 dataset, the system achieved a peak signal-to-noise ratio of up to 54 dB and a structural similarity index of 92%. These results indicate that the proposed model delivers high accuracy and strong diversity in painting art design. It effectively solves current problems in segmentation precision and style fusion, offering new ideas for artistic creation and supporting the development of intelligent painting systems.

Povzetek: Študija predstavlja izboljšan model za generiranje umetniških slik, ki omogoča natančnejšo obdelavo in bolj naravno združevanje slogov ter izboljšuje kakovost ustvarjenih slik.

1 Introduction

With the rapid development of technology, the field of painting art is undergoing major changes through advanced artificial intelligence and graphics processing. These technologies allow creators to explore diverse styles efficiently and offer new ways to innovate, reshaping the art landscape [1]. artistic image generation systems reduce the difficulty of creation, improve efficiency, and encourage public participation, bringing new energy to the art world [2]. Therefore, designing an intelligent artistic image generation system is of great value to today's society. At present, the main approaches to system design fall into two categories: rule-based and data-driven. Rule-based methods rely on manually defined rules, making them less flexible and unable to meet complex artistic needs. Data-driven methods can learn patterns from data, but their performance is easily affected by data quality, and they often struggle to understand deep semantic meaning in art [3]. A system that understands deep semantics and supports diverse styles is urgently needed. Most artistic image generation systems focus on semantic segmentation and style transfer. Convolutional Neural Network (CNN) extracts multi-scale features, helping the model adapt to different object sizes and scene details, which improves its ability to model semantic information in images [4]. Generative Adversarial Network (GAN) learns the distribution of various styles and generates images that are diverse, natural, and realistic [5]. Based on this, the paper combines CNN and GAN with an interactive GrabCut to build a artistic image generation system. This system aims to enhance the creativity of artworks and support the spread and inheritance of painting culture. The research aims to address the issue of insufficient semantic segmentation accuracy in existing artistic image generation systems. By integrating improved CNN and GrabCut algorithms, the goal is to achieve segmentation performance with a Mean Intersection over Union (mIoU) of $\geq$ 90% on the PASCAL VOC dataset. Secondly, regarding the issue of naturalness in style transfer, the semantic guided GAN architecture can achieve a

Structural Similarity Index Measure (SSIM) of ≥ 0.9 for the generated images. Finally, regarding the practicality of the system, set real-time processing performance targets: inference latency ≤ 62ms and memory usage ≤ 800MB at 1080p resolution.

2 Related works

Semantic segmentation and style transfer technologies had their own strengths and helped solve complex problems across various fields. Scholars in China and abroad conducted related studies. For example, Xie B et al. noticed that existing methods in domain adaptive semantic segmentation often ignored the internal connections within training data, making it difficult to handle cross-domain semantic changes. They put forward a semantic-guided pixel contrastive single-stage adaptation framework. By studying centroid-aware pixel contrast, the method significantly improved segmentation results in cross-domain tasks [6]. To address the problem of feature loss in tunnel lining crack detection, Zhou Z et al. introduced the SCDeepLab algorithm under the DeepLabv3+ encoder-decoder framework. The results showed that SCDeepLab reached a mean intersection over union of 77.41% and a mean pixel accuracy of 84.42 % on their custom dataset [7]. To improve the low classification accuracy in early cancer tumor detection, Samudrala S et al. built a hybrid semantic segmentation network. They combined the DenseNet-121 model with an attention-based pyramid scene parsing network and added an attention gating mechanism to enhance feature quality. Experimental results indicated a prediction accuracy of 94.68% [8]. To solve the low accuracy in embedding and extracting secret information from steganographic images, Garg M et al. proposed a hybrid steganographic technique based on neural style transfer and Generative Adversarial Network. Their results showed a peak signal-to-noise ratio of 44.175 dB, a structural similarity index of 0.9958, and a visual information fidelity score of 0.954 [9]. To handle the challenge of cross-language style transfer in automatic font synthesis, Li C et al. put forward a model that captured font style features using multi-level attention. Their results demonstrated strong generation performance in cross-language font style transfer tasks [10].

In the field of painting, both theoretical and practical applications had matured in some areas, and many researchers from different countries had explored them in depth. For instance, to solve the lack of user interaction in traditional Chinese painting and the difficulty of transforming it into realistic images, Chung C Y et al. combined cycle-consistent Generative Adversarial Network and Pix2Pix with a tag function and border-enhanced GAN. The generated images were more realistic and provided users with a novel artistic experience, which could also serve as a reference for other fuzzy boundary image transformations [11]. To address the difficulty of classifying specific types of paintings, Ugail H et al. proposed an algorithm that used transfer learning in deep neural networks for feature extraction. They combined this with a support vector machine binary classifier and integrated edge detection. The method achieved a classification accuracy of 98%, supporting the use of visual analysis in artwork identification tasks [12]. To solve the lack of quantitative methods for evaluating the preservation status of cultural painting heritage, Eom T H et al. used image analysis techniques. They analyzed image color spaces and calculated the shape and area of the damage. Their results confirmed that digital color information could objectively distinguish damaged areas and help assess preservation conditions [13]. To tackle the challenge of generating emotional captions in paintings, Lu Y et al. developed a new model that included facial expressions and human pose features. These were combined with commonly used object functions to generate emotional captions for abstract visual artworks [14]. To explore the use of terahertz time-domain imaging in painting research, Fukunaga K et al. applied this method to perform point-by-point scans of samples. They combined spectral and two-dimensional imaging and used the signal reflections from different layers to obtain stratigraphic information and even tomographic images. This made it possible to perform non-destructive inspections of underpainting conditions [15]. The summary table of the relevant works mentioned above is shown in Table 1.

Table 1: Table of related works

Author	Method	Dataset/Application Scenarios	Result	Main limitations
Xie et al.	Semantic guided pixel contrast domain adaptation framework	Cross domain segmentation of urban landscape	mIoU: 76.2%	High computational complexity (8.3 FPS)
Zhou et al.	Swin CNN hybrid tunnel crack detection	Tunnel lining crack dataset	mPA: 84.42%	Edge mIoU decreased by 12%
Samudrala	DenseNet-121+Attention	Pathological image of breast cancer	Accuracy: 94.68%	Only applicable to medical images

	PSPNet			
Garg et al.	Neural style steganography	COCO-Style	PSNR: 44.175 dB	Single style transfer
Li et al.	Cross language font style transfer	Multilingual Font Dataset	User preference: 82%	Limited style diversity
Chung et al.	Boundary enhanced GAN Chinese painting conversion	Ink Painting Dataset	Realistic rating: 4.1/5.0	Lack of semantic understanding
Ugail et al.	Deep Transfer Learning for Painting Classification	WikiArt	Classification accuracy: 98%	No ability to generate
Eom et al.	Digital Diagnosis of Painting Heritage	Cultural heritage painting	Damage recognition rate: 89%	Only analyze without repair function
Lu et al.	Emotional Description Generation Model	Abstract artwork	Emotional compatibility: 0.81	Disconnected from the generation process
Fukunaga	Terahertz imaging tomography analysis of painting	Multi layer oil painting samples	Tomography resolution: 0.1mm	High device dependency

Based on the above content, it can be concluded that although significant achievements have been made in the field of art image processing, the development of intelligent artistic image generation systems still faces several key challenges. For example, current methods generally lack a comprehensive grasp of the entire process of artistic creation, and are limited to basic object recognition in semantic understanding, making it difficult to capture the deep compositional rules and aesthetic principles contained in painting works; In terms of style processing, most adopt a globally unified transfer strategy, which cannot achieve precise adaptation to the semantic structure of the image. Therefore, the proposed model combining semantic segmentation and style transfer offered good practicality. It aimed to enhance creative efficiency and support personalized design in the context of artistic innovation, meeting the diverse needs of artistic creation.

3 Semantic segmentation and style transfer technologies for artistic image generation system design

3.1 Optimization of semantic segmentation using improved CNN-based GrabCut

In traditional image semantic segmentation algorithms, one of the most well-known methods is the interactive image segmentation algorithm GrabCut [16]. Its advantages lie in high segmentation accuracy, the ability to handle complex image backgrounds and foreground objects, and the low requirement for user expertise, as it only needs simple user interactions. In artistic image generation system design, it is usually necessary to classify each pixel in the image into different semantic categories and identify regions corresponding to objects such as people or scenery, laying the foundation for artistic processing. Therefore, this study applies GrabCut to segment and classify images semantically, and then builds the artistic image generation system based on it. During the design process, GrabCut first initializes the foreground and background based on pixel information, and then performs iterative optimization, as shown in Figure 1.

As shown in Figure 1, GrabCut first reads the input data. It then marks a rectangle based on the user's input, where the inside is considered foreground and the outside is background by default. Next, an initialized Gaussian Mixture Model (GMM) is used to model the foreground and background separately. GMM fits pixel color distributions using multiple Gaussian distributions. It calculates parameters such as the mean and covariance of each Gaussian distribution based on many pixel samples, which accurately describe the color features of the foreground and background [17]. Then, an energy function is constructed with data terms and smoothness terms.

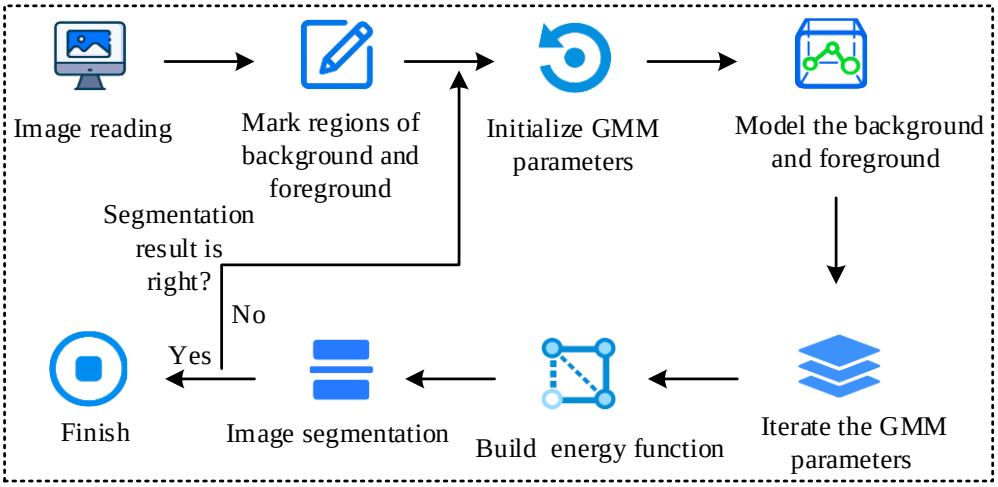


Figure 1: GrabCut interactive image segmentation algorithm operation flow chart

The data terms measure the matching degree between pixels and GMM, while the smoothness terms ensure continuity of the segmentation boundary. Finally, the GMM parameters and energy function are updated iteratively, and graph cut is performed until the convergence condition is met to output the result. The calculation process of the GMM fitting model is shown in Equation (1).

$$P(y|\theta)=\sum_{k=1}^{K}\alpha_k\phi(y|\theta_k) \tag{1}$$

In Equation (1), α_k represents a coefficient that satisfies $\alpha_k \geq 0$ and $\sum_{k=1}^{K}\alpha_k=1$. $\phi(y|\theta_k)$ denotes the Gaussian distribution density. The value range of the Gaussian distribution density is shown in Equation (2).

$$\theta_k=(\mu_k,\sigma_k^2) \tag{2}$$

In Equation (2), k represents the k-th sub-model. However, since most of the input images in practice are red, green, and blue channel images, each sub-model in the GMM is a 3D Gaussian distribution. The multivariate Gaussian distribution is shown in Equation (3).

$$N(\bar{x}|\bar{u},\sum)=\frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\sum|^{\frac{1}{2}}}\exp[-\frac{1}{2}(\bar{x}-\bar{u})^T\sum^{-1}(\bar{x}-\bar{u})] \tag{3}$$

In Equation (3), $\bar{x}$ denotes a vector of dimension D, and $\bar{u}$ represents the mean value of these vectors. Based on the Gaussian model, the GrabCut method can accurately capture object boundaries and model the object in the image, achieving good segmentation performance. However, the artistic image generation system often contains a large amount of pixel information and long-range semantic features. GrabCut has high computational complexity, especially when handling high-resolution images. It requires heavy computation, takes a long time, has poor practicality, and struggles to capture complex features, which reduces segmentation accuracy. Therefore, a method is needed to improve its efficiency and accuracy. CNN is a deep learning model whose convolutional layers can automatically learn multi-level image features, from low-level textures to high-level semantics. Its strong nonlinear expression ability can fit complex segmentation boundaries and adapt to various complex scenes, greatly improving segmentation accuracy and efficiency [18]. However, traditional CNNs still have limitations in handling long-range dependencies and global information. The Convolutional Block Attention Module (CBAM) helps CNN automatically focus on important regions and features, improving its ability to capture global information. In addition, compared to the SE module that only focuses on the computational complexity of channel dimension attention or self attention mechanisms, CBAM achieves cross channel feature recalibration and cross spatial position perception at a lower computational cost through cascaded spatial attention and channel attention modules. The dual path design of CBAM can effectively handle two common key features in art works. Its basic principle is as follows: the channel attention path focuses on channel sensitive features such as pigment color distribution, which is crucial for tone control in style transfer; The spatial attention path strengthens the modeling of spatial relationships such as stroke direction and composition layout, which plays a decisive role in maintaining the structural integrity of the artwork. Therefore, this study introduces CBAM to improve CNN and then integrates the improved CNN into the traditional segmentation algorithm. The structure of the improved CNN is shown in Figure 2.

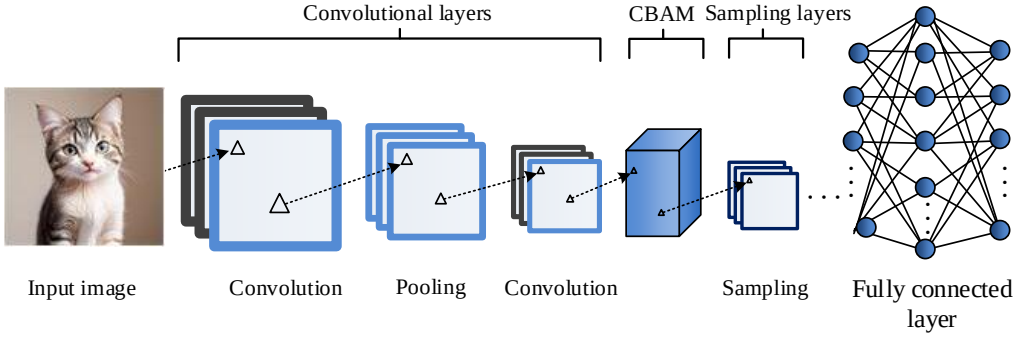


Figure 2: CBAM improved CNN structure diagram

As shown in Figure 2, the target image is first input. Then, the convolution kernels in the convolutional layers slide over the image to perform convolution operations, extracting different features such as edges and textures. Multiple convolutional layers extract features at different levels. The pooling layer performs downsampling on the output of the convolutional layer to reduce data volume while retaining key features. After another convolution operation, CBAM captures long-range dependencies to obtain feature representations. Next, sampling is performed through operations such as deconvolution or transposed convolution, restoring the feature map to a size close to the original image. Skip connections are added to fuse features at different levels and improve segmentation accuracy. Finally, a classifier such as Softmax in the fully connected layer classifies each pixel and determines its category. The convolution operation is calculated as shown in Equation (4).

$$Y_k = f(W_k * x) \tag{4}$$

In Equation (4), x represents the input value, which interacts with the convolution kernel W_k connected to the k-th feature. The result is processed by a nonlinear activation function $f()$. After convolution, data is reduced in the pooling layer to increase the receptive field. This process is shown in Equation (5).

$$Y_{ijk} = \max_{(p,q)\in\Re_v} x_{kpq} \tag{5}$$

In Equation (5), x and the k-th feature map are merged using a specific method, and the final result is represented by Y_{ijk}. The element located at coordinate position (p,q) in the merging region is marked as x_{kpq}, and $\Re_v$ denotes the receptive field. After the convolution and pooling operations, classification is performed through a fully connected layer. The commonly used loss function is shown in Equation (6).

$$E = \frac{1}{N}\sum_{n=1}^{N}\log(p_{nk}), k \in [0,1\ldots,K-1] \tag{6}$$

In Equation (6), k represents the number of classification categories, and p_{nk} indicates the probability value of a classification category. The improved CNN is used to optimize the traditional semantic segmentation algorithm through fusion. The process is shown in Figure 3.

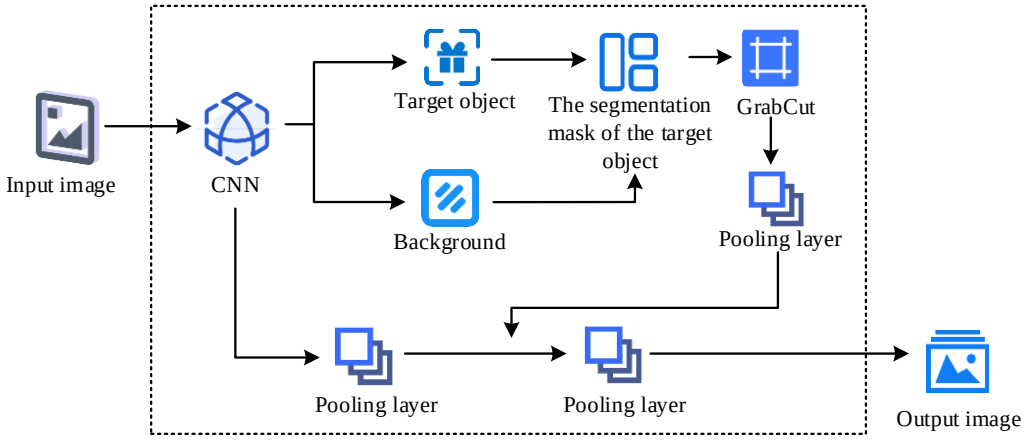


Figure 3: Improved CNN-GrabCut model structure

Figure 3 shows that the system adopts a hierarchical feature fusion mechanism, and the deep features extracted by the CNN network are first used to generate initial segmentation suggestions, which serve as input regions for the GrabCut algorithm. GrabCut outputs a refined segmentation mask through iterative energy

minimization, which is then converted into a weight matrix that matches the CNN feature map, rather than directly performing feature pooling operations. Specifically, the multi-scale features extracted by CNN are transmitted through horizontal connections; The segmentation mask output by GrabCut is processed through differentiable binarization to generate spatial attention weights. The weight matrix is point multiplied with the shallow feature map of CNN to enhance feature selectivity, rather than traditional pooling operations. This design not only retains the advantage of precise boundary segmentation in GrabCut, but also achieves the fusion of contextual information through CNN features. The horizontal pooling is calculated as shown in Equation (7).

$$y_{cj}^{h} = \frac{1}{W} \sum_{0 \le i \le w} X_{c,i,j} \tag{7}$$

In Equation (7), c represents the number of channels. W denotes the width of the feature map in the horizontal dimension. i refers to the i-th row in the feature map structure, and j refers to the j-th column. After horizontal pooling, vertical pooling is performed as shown in Equation (8).

$$y_{c,j}^{v} = \frac{1}{H} \sum_{0 \le i \le H} X_{c,i,j} \tag{8}$$

In Equation (8), H represents the height of the feature map. The data features obtained through horizontal and vertical pooling are fused, and the corresponding is shown in Equation (9).

$$y = Scale(x, \sigma(f(y^{h}))) \tag{9}$$

In Equation (9), $Scale()$ denotes the element-wise multiplication, σ represents the Sigmoid activation function, and f represents the convolution.

3.2 Construction of artistic image generation system combining improved CNN-GrabCut and GAN

Although the CNN-GrabCut model combines the strengths of improved CNN and GrabCut and can efficiently and accurately process image information through convolution and iteration, it lacks adaptability to diverse styles required in artistic image generation system design. Its performance is limited by the style of the pre-trained model and cannot precisely capture style features. Therefore, a style transfer technique is introduced to improve it [19]. In terms of model integration process, the system adopts a strict temporal control data transmission mechanism: the multi-scale mask output by the semantic segmentation network will be tensor concatenated with the 128-dimensional style vector extracted by the style encoder in the third downsampling stage. This key integration point is validated through a timestamp marked log system to ensure feature fusion is completed before the fourth residual block of the GAN generator. Specifically, after the semantic segmentation network completes forward propagation, the system will trigger the following sequence operations: firstly, perform bilinear upsampling on the highest layer semantic mask at t+2ms; The second is to perform Hadamard product operation with style features at t+5ms; Thirdly, channel dimension alignment is achieved through 1 × 1 convolution at t+7ms. This process is implemented asynchronously and parallelly through CUDA event flow, ensuring temporal traceability. A GAN improved with residual connections shows strong adaptability and high efficiency in style transfer. Its complete structure is shown in Figure 4.

As shown in Figure 4, the improved GAN includes a generator, residual connections, and a discriminator. The generator takes a random vector as input and learns the underlying data distribution to map it into fake samples. Residual connections solve problems such as gradient vanishing and degradation before inputting the data into the discriminator. The discriminator receives samples from both the real data distribution and the generator and outputs a probability indicating whether a sample comes from real data.

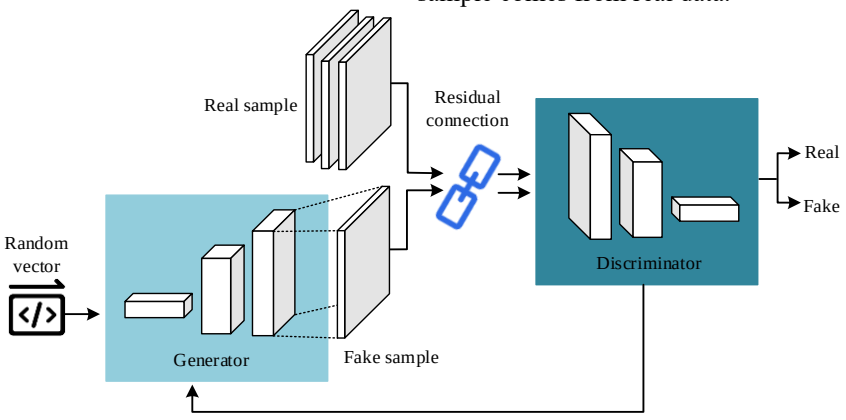


Figure 4: Improved GAN structure diagram

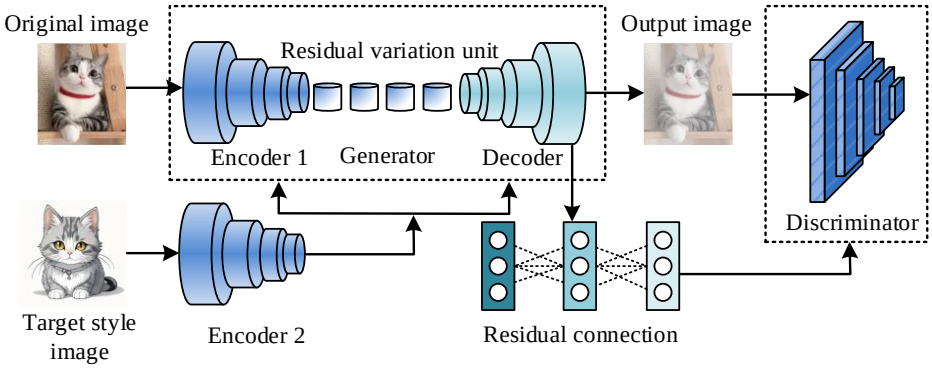


Figure 5: Workflow of style transfer technology based on improved GAN

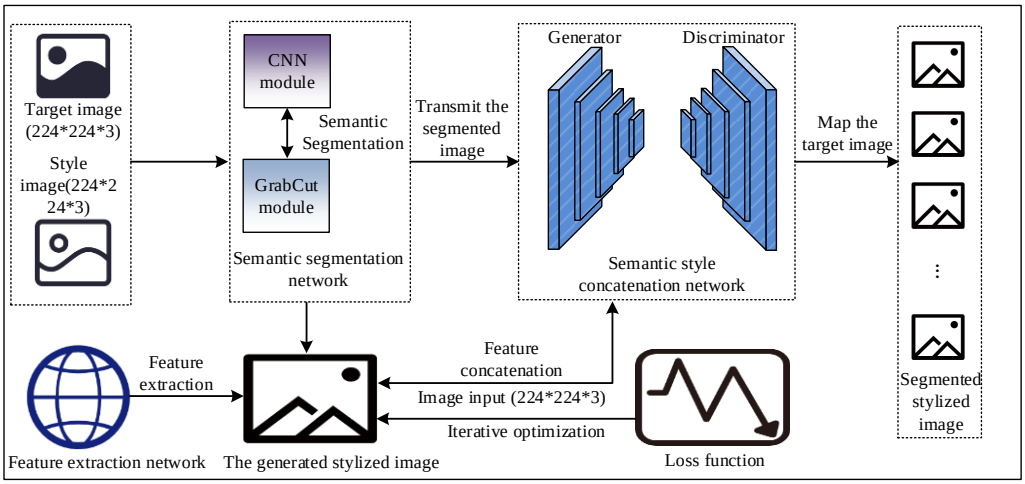


Figure 6: GCG artistic image generation system structure diagram

Through adversarial training, the generator minimizes the probability that the discriminator correctly identifies fake samples, while the discriminator maximizes its accuracy until reaching Nash equilibrium. The loss function of the generator is shown in Equation (10).

$$L_G = H(1, D(G(z))) \tag{10}$$

In Equation (10), G denotes the generator, D the discriminator, H the cross-entropy, z the random data, and $D(G(z)$ the probability judgment of generated data. The loss function of the discriminator is shown in Equation (11).

$$L_D = H(1, D(x))H(0, D(G(z))) \tag{11}$$

In Equation (11), x indicates real data, $H(1, D(x))$ the distance between real data and 1, and $H(0, D(G(z)))$ the distance between generated data and 0. Through adversarial training between the generator and discriminator, the generator learns to embed source style features into images and learns the target style distribution to achieve style transfer. The model based on the improved GAN for style transfer is shown in Figure 5.

As shown in Figure 5, the model includes a generator network, a discriminator network, a style encoder, and residual connections. The generator network adopts an encoder, residual transformation units, and a decoder. During image transformation, the original content image and the target style image are encoded separately. The style encoding is transformed to obtain a condition vector. The latent variables and the condition vector are combined and processed by the residual transformation and decoder to generate a new stylized image. The decoder's computation is shown in Equation (12).

$$L_s = E_{X_s, Z} \left\| s_0 - \varphi(G(z, s_0)) \right\|_2^2 \tag{12}$$

In Equation (12), E represents encoder 1, φ represents style encoder 2, s_0 denotes the latent space features generated by style encoder 2, and z denotes the latent space features generated by encoder 1 [20]. The formula for calculating the transformation from the original image to the stylized image using the style encoder is shown in Equation (13).

$$z = E(x_c), \varphi(x_s) \to s_0, G(z, \varphi(x_s)) \to x_g \tag{13}$$

In Equation (13), x_g denotes the generated stylized image, φ the encoder, and s_0 the predicted style condition. This mechanism enables efficient and accurate style transfer. Finally, the improved GAN-based style transfer model is combined with the improved CNN-GrabCut model to build the GCG artistic image generation system. The system structure is shown in Figure 6.

Figure 6 shows that the original image is first input into a semantic segmentation network for semantic mask segmentation and downsampling processing; Meanwhile, the style image is directly input into the semantic style fusion network for style feature encoding. The outputs of the two paths are combined in a feature fusion network, iteratively optimized through content loss and style loss calculations, and finally generated into the target output image by the feature extraction network. The downsampling operation applied to the semantic mask of the original image is calculated as shown in Equation (14) [21-22].

$$m_l^s = downsampling(m, scale(l)) \tag{14}$$

In Equation (14), s in m_l^s represents the semantic mask from the first layer of the model structure. The downsampling ratio is denoted by $scale(l)$. m is the intermediate feature map extracted from the semantic segmentation network, which will be used to generate semantic masks in the future. After the downsampling is completed, it is important to note that the resolution of the mask becomes consistent with the resolution of the style features in each layer of the model. Once this consistency is achieved, the next step is to extract features from the mask, as shown in Equation (15).

$$s_l^e = f(x_t) \tag{15}$$

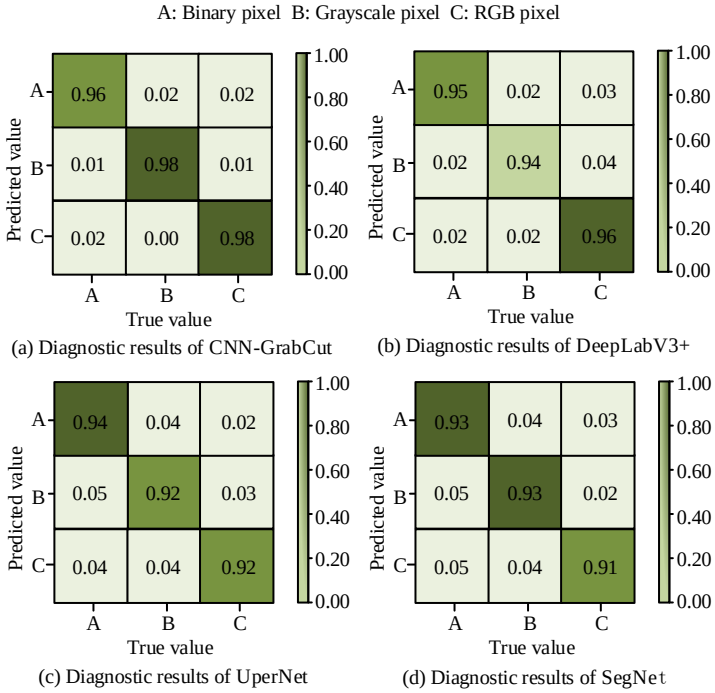
In Equation (15), s_l^e represents the first layer style feature in the GAN style encoder, f represents the style feature extraction network, and x_t represents the target image. Based on the above content, it can be concluded that the proposed GCG system adopts a region adaptive style transfer strategy, which achieves deep coupling between semantic segmentation and style transfer at the architecture level. After the system generates pixel level semantic masks through an improved CNN GrabCut module, the masks will serve as spatial weight matrices to guide the style transfer process, maintaining global style consistency while achieving local style adaptation. Specifically, the multi-scale masks output by the semantic segmentation network will undergo tensor multiplication with the feature maps extracted by the style encoder, resulting in differences in style transfer intensity between different semantic regions. For foreground subject areas such as characters or buildings, the system adopts a strong style transfer coefficient to highlight artistic expression; For background areas such as the sky or water surface, a weak style transfer coefficient should be applied to maintain a natural transition. This region adaptation mechanism achieves end-to-end training through differentiable rendering. When calculating style loss, the semantic segmentation results of the content image will automatically establish corresponding relationships with similar regions of the style image, ensuring that texture features are transmitted between semantically equivalent regions.

4 Validation of the artistic image generation system based on semantic segmentation and style transfer

4.1 Validation of the CNN-GrabCut algorithm

To evaluate the performance of the CNN-GrabCut algorithm in semantic segmentation, it was compared with the DeepLabV3+, UperNet, and SegNet algorithms. The operating system was Linux, the deep learning framework was Keras, the optimizer was SGD, and the programming language was Python 3.10.12. The experiments were conducted on a system equipped with an NVIDIA RTX 3080 GPU and 32GB of memory. The PASCAL VOC dataset was used for both training and testing. All images were uniformly adjusted to 224 × 224 pixels during the preprocessing stage. In order to enhance the generalization ability of the model, data augmentation techniques were used during the training process, including random horizontal flipping, random cropping, and color jitter. The experimental parameters are set as follows: the model is trained using a stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.001, momentum is set to 0.9, and the learning rate decays by 0.1 every 10 epochs. The batch size for training is 32, and the total training time is 50 epochs. The performance of the model is evaluated through metrics such as pixel accuracy, Mean Intersection over Union (MIoU), Mean Pixel Accuracy (MPA), F1 score, and Area Under the ROC Curve (AUC). CNN-GrabCut, DeepLabV3+, UperNet, and SegNet were tested on pixel accuracy across three-pixel types: binary, grayscale, and RGB. The results are shown in Figure 7.

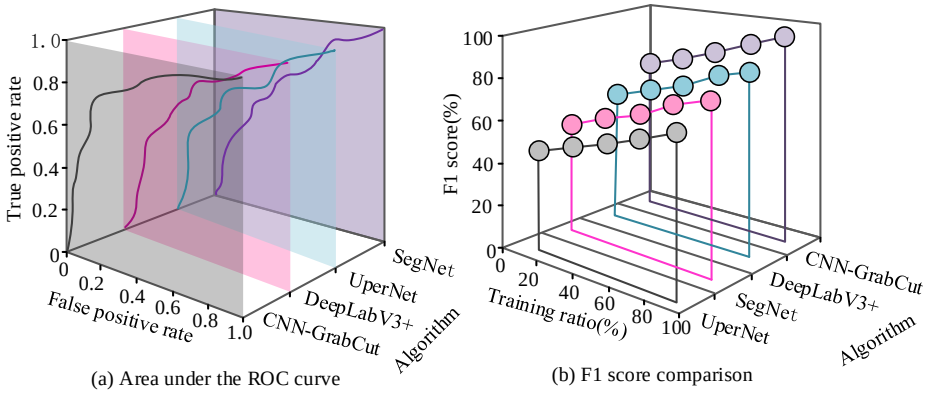
As shown in Figure 7(a), the CNN-GrabCut algorithm achieved a pixel accuracy of 98% for grayscale pixels, 96% for binary pixels, and 98% for RGB pixels. The highest accuracy achieved by DeepLabV3+ was 96%. The maximum accuracy of UperNet and SegNet reached 94% and 93%, respectively. These results indicated that CNN-GrabCut achieved better pixel accuracy. To further evaluate the performance of each algorithm, the study compared the Area Under Curve (AUC) value and the F1 score, as shown in Figure 8.



(a) Diagnostic results of CNN-GrabCut (b) Diagnostic results of DeepLabV3+

(c) Diagnostic results of UperNet (d) Diagnostic results of SegNet

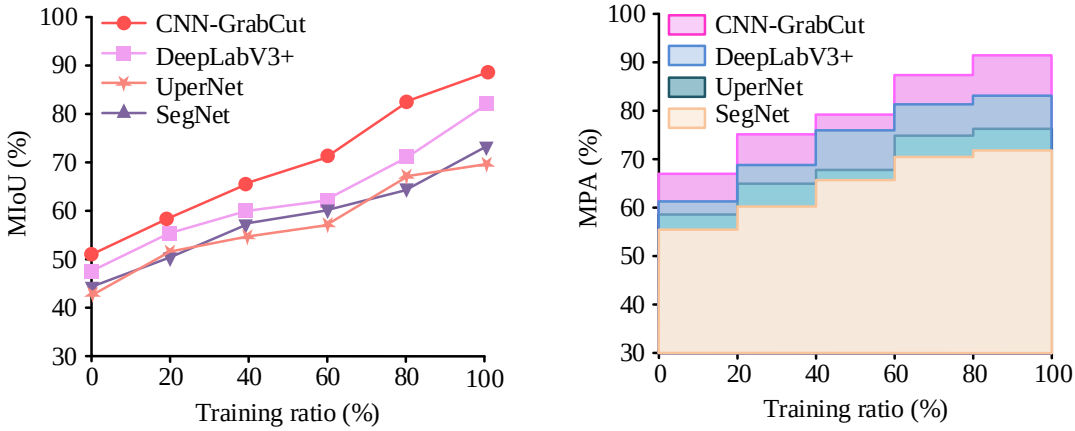
Figure 7: Pixel accuracy experimental results



(a) Area under the ROC curve (b) F1 score comparison

Figure 8: Comparison of AUC value and F1 score

Figure 8(a) shows that the ROC curve of CNN-GrabCut was closest to the top-left corner, with an AUC of 0.873, which was significantly higher than the comparison models, which reached 0.821, 0.683, and 0.623, respectively. This demonstrated better classification performance and prediction reliability. As shown in Figure 8(b), the F1 score continuously increased with more training epochs. The lowest F1 score of CNN-GrabCut was 80%, which was notably higher than the other models, further confirming its strong classification ability. Overall, CNN-GrabCut outperformed the comparison models in both classification prediction and recall rate. To further validate the segmentation performance, the study compared the models using Mean Intersection over Union (MIoU) and Mean Pixel Accuracy (MPA), as shown in Figure 9.



(a) MIoU comparison chart (b) MPA comparison chart

Figure 9: Comparison results of MIoU and MPA

Table 2: Statistical testing and robustness results

Project		Index	CNN-GrabCut	DeepLabV3+	UperNet	SegNet
Statistical test results		MIoU mean (%)	91.2	82.7	66.8	72.5
		MIoU confidence interval question (%)	[89.9,92.5]	[80.6,84.8]	[63.4, 70.2]	[69.7, 75.3]
		SSIM mean	0.918	0.835	0.775	0.820
		SSIM confidence interval question	[0.910,0.926]	[0.823,0.847]	[0.760,0.790]	[0.810,0.830]
		LPIPS	0.15	0.23	0.31	0.29
Robustness testing	Gaussian noise	△MIoU (%)	[-4.1,-2.3]	[-8.7,-6.3]	[-13.5,-10.7]	[-10.2,-8.4]
		△SSIM	[-0.05,-0.03]	[-0.10,-0.08]	[-0.16,-0.14]	[-0.13,-0.11]
	Occlusion condition	△MIoU (%)	[-6.0,-4.2]	[-12.5,-10.1]	[-19.1,-16.3]	[-15.6,-13.2]
		△SSIM	[-0.08,-0.06]	[-0.15,-0.13]	[-0.22,-0.20]	[-0.19,-0.17]

As shown in Figure 9(a), with the increase in the training data ratio, the MIoU of CNN-GrabCut remained higher than those of the other models. Its highest MIoU reached 92%, significantly exceeding the values of DeepLabV3+ (83.2%), UperNet (67.5%), and SegNet (73.2%). Figure 9(b) shows that the MPA of CNN-GrabCut reached up to 92.1% with full training, consistently outperforming the other models throughout the training process. These results indicated that the proposed CNN-GrabCut algorithm did not suffer from overfitting or inappropriate learning rate settings and achieved better segmentation performance, offering strong support for accurate segmentation in the design of artistic image generation systems. To more comprehensively reflect the generalization ability and robustness of the model, the study uses a five-fold cross-validation method for testing, and simultaneously sets two test conditions: a Gaussian noise of 0.1 and a 20% occlusion area. From this, the statistical test and robustness results can be obtained, as shown in Table 2.

It can be seen from Table 2 that the CNNS-GRABCut method shows significant advantages in statistical performance and robustness tests. Its average MIoU reaches 91.2% and the confidence interval range is the narrowest, only 2.6%. This is attributed to the model integrating the advantages of the CBAM attention mechanism and the GrabCut algorithm. By effectively suppressing background interference through dual attention of channels and Spaces, and simultaneously combining horizontal and vertical feature pooling to retain precise boundary information, the consistency of feature extraction and the stability of segmentation are enhanced. In terms of style transfer, the average SSIM is 0.918, close to the theoretical maximum value, indicating that the semantically guided GAN architecture achieves precise adaptation of style features in the target area through the interactive operation of the style encoder and the semantic mask, avoiding excessive distortion of the background. Facing the interference of Gaussian noise and regional occlusion, CNN-GrabCut demonstrates outstanding robustness, with the MIoU decrease controlled within 4.1% and the SSIM decrease not exceeding 0.08. This is attributed to the adaptability of the Gaussian mixture model to the noise distribution and

the anti-interference characteristics of the residual connection. Meanwhile, the multi-scale feature fusion and semantic region adaptive mechanism further enhance the model's fault tolerance for local information loss. In addition, the style transfer design of hierarchical feature fusion and timing control not only optimizes boundary accuracy and context understanding, but also achieves real-time performance in high-resolution images through asynchronous parallel processing. Although the computing resource requirements are slightly higher, the model's comprehensive performance in terms of accuracy and stability makes it an ideal choice for generating artistic images in complex scenarios. Furthermore, the study used Learned Perceptual Image Patch Similarity (LPIPS) for analysis, which assesses perceptual quality by comparing the similarity of images in the feature space. It can be found that the LPIPS value of the GCG system is 0.15, indicating that it is superior to other systems in terms of perceived quality. Finally, to investigate the performance contribution of each module to the overall system, ablation experiments were conducted, and the results are shown in Table 3.

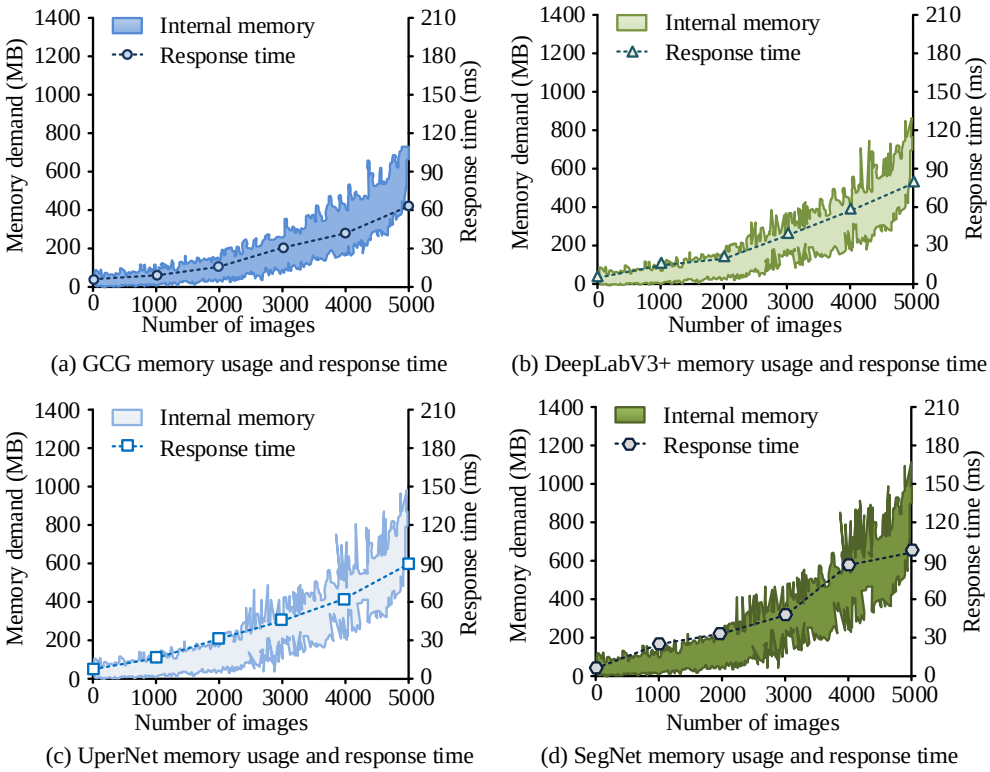
Table 3: Results of ablation experiment

Evaluation	Baseline Model (CNN)	+ GrabCut	+ N63CBAM Attention	+ Semantic guided GAN	Complete model
MIoU (%)	82.1	86.7	88.9	90.4	91.2
SSIM	0.801	0.842	0.883	0.908	0.918
LPIPS	0.185	0.153	0.124	0.105	0.098
Memory usage (MB)	580	620	650	695	710
Single frame latency (ms)	45	53	58	61	62
Throughput (FPS)	22.2	18.9	17.2	16.4	16.1
GPU utilization rate (%)	68	72	75	78	80

Table 3 shows that the basic model performs the best in real-time, achieving a throughput of 22.2 FPS with a single frame processing delay of 45ms. However, the quality of semantic segmentation and style transfer is significantly lagging behind. After introducing the GrabCut algorithm, the latency increased to 53ms and the throughput decreased to 18.9 FPS. This is due to the additional computational load brought by the GMM iterative optimization process, but the 4.6% improvement in MIoU verified its boundary optimization effect. The addition of CBAM attention module further increases the latency to 58ms, but the GPU utilization rate increases to 75%, indicating that its channel and spatial attention mechanism effectively utilizes parallel computing resources. At the same time, SSIM is increased from 0.842 to 0.883, proving the key role of feature enhancement in style fidelity. Although semantic guided GAN introduces a maximum latency overhead of 61m, it achieves style semantic alignment through matrix multiplication, causing LPIPS to plummet from 0.124 to 0.105, and GPU utilization to reach 78%, reflecting the efficient utilization of computing resources by residual connections and conditional vector fusion. The complete model maintains a real-time processing capability of 16.1 FPS with a latency of 62ms, while the MIoU reaches 91.2%, SSIM 0.918, The balance between performance and efficiency is demonstrated, and the dot multiplication operation of multi-scale masks and style features is implemented asynchronously in parallel through CUDA event streams. With only an additional 1ms delay, it achieves an additional 1.3% MIoU improvement compared to a single GAN variant. This confirms the rationality of the design in terms of computing resource allocation. The above results are due to the contribution of residual connections to style transfer. Firstly, the original content features and style condition vectors passed across layers in the fourth residual block of the generator form a dynamic gating mechanism, which suppresses inconsistent style noise. The second is that the residual skip connection of the decoder recombines shallow spatial details with deep semantic features.

4.2 Evaluation of the system based on semantic segmentation and style transfer

After validating the performance of CNN-GrabCut, the study further evaluated the performance of the GCG artistic image generation system by comparing it with systems built using DeepLabV3+, UperNet, and SegNet. The experiments were conducted on a workstation with a high-performance GPU, using the Windows operating system. The dataset used was ArtBench-10, a well-known dataset covering diverse styles and genres. The memory usage and response time of the four systems when processing different amounts of images are shown in Figure 10.



(a) GCG memory usage and response time

(b) DeepLabV3+ memory usage and response time

(c) UperNet memory usage and response time

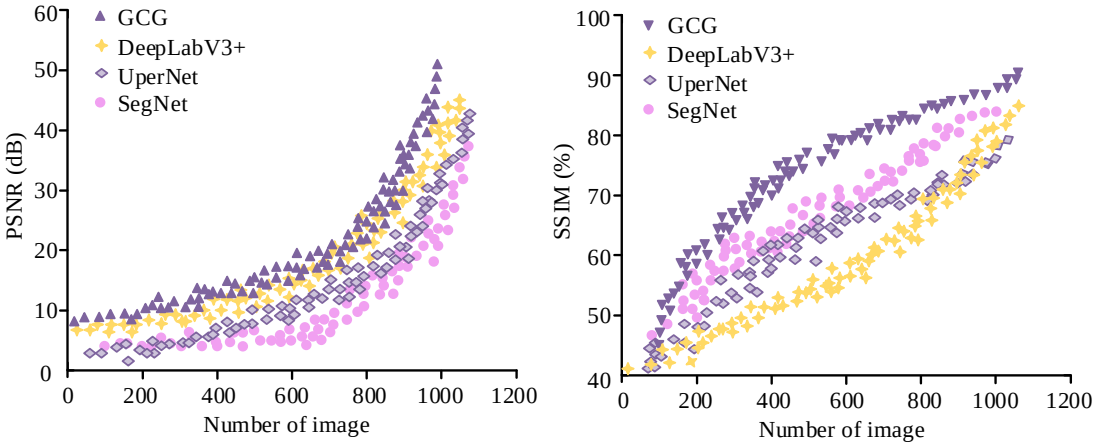
(d) SegNet memory usage and response time

Figure 10: Experimental results of system memory and response time

As shown in Figure 10(a), when the number of processed images reached 5000, the GCG system used 710 MB of memory with a response time of 62 ms. Figure 10(b) shows that the DeepLabV3+ system had the highest memory usage of 880 MB and the longest response time of 81 ms. Figures 10(c) and 10(d) indicates that the UperNet and SegNet systems had maximum response times of 90 ms and 98 ms, and maximum memory usage of 1000 MB and 1100 MB, respectively. These results demonstrated that the GCG system had the shortest response time and the lowest memory usage when processing large volumes of images, indicating a more efficient image processing capability. To further assess image quality, the study compared the Peak Signal-to-Noise Ratio (PSNR) and SSIM of the four systems, as shown in Figure 11.

As shown in Figure 11(a), with an increasing number of image samples, the GCG system achieved a maximum PSNR of 54 dB, which was higher than the comparison systems at 46 dB, 44 dB, and 38 dB. Figure 11(b) shows that the GCG system reached a maximum SSIM of 92%, while the maximum SSIM values for the DeepLabV3+, UperNet, and SegNet systems were 84%, 78%, and 83%, respectively. These results indicated that the GCG system produced images with the highest structural similarity to the original, achieving the best overall image quality. Finally, in order to further demonstrate the performance of the artistic image generation systems of the four systems, the study compared the image diversity of the four systems. The ArtBench-10 test set was used for the experiment, and in response to the limitations of traditional Inception Score and FID indicators, which are based on natural image classifiers and difficult to accurately capture the unique dimensions of artistic styles, a professional art evaluation mechanism was introduced. Five professional painters rated 200 sampled works for style originality, using a double-blind testing method and a 5-point scale. The result can be obtained as shown in Figure 12.

As shown in Figure 12, the GCG system reached a maximum image count of 4000 and the highest number of categories at 400. Among the comparison systems, the DeepLabV3+ system performed the best, with a maximum of 3500 images and 380 categories. In addition, according to the calculation research method, the average score is 4.2 points, significantly higher than the baseline model's 3.1 points. 63% of generated works are considered to have expanded traditional forms of expression through innovative combinations of color rhythm and stroke tension while maintaining the typical characteristics of the genre. The above may be due to the artistic value derived from the implicit modeling ability of the system for emotional features. By analyzing the color combinations and composition rules in a large number of classic works, the system automatically learns the complex mapping relationship between emotional expression and visual elements. These results indicate that the diversity of works generated by the system can effectively stimulate the inspiration of art creators.



(a) Results of different systems in PSNR (b) Results of different systems in SSIM

Figure 11: Experimental results of PSNR and SSIM

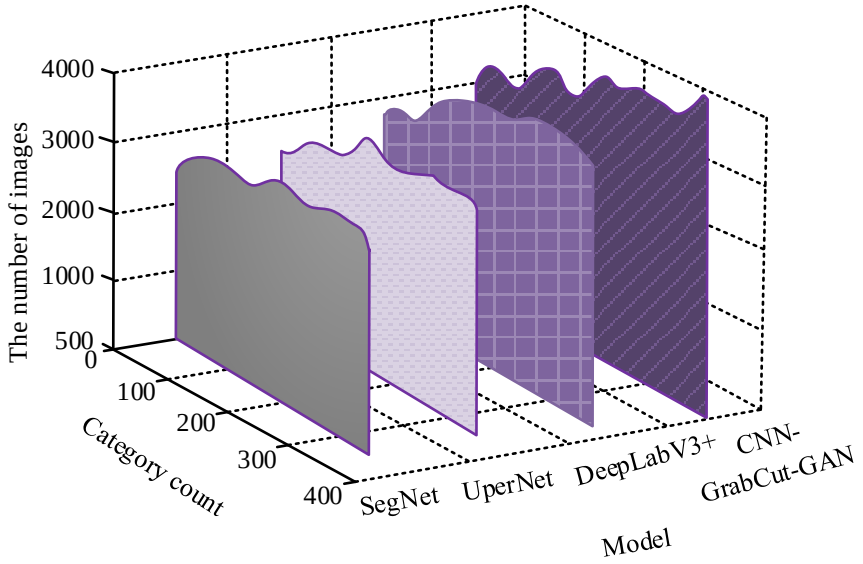


Figure 12: Image diversity experiment results

5 Discussion

A GCG painting art image generation system has been proposed, which combines the improved CNN GrabCut algorithm with enhanced GAN to compare semantic segmentation and style transfer with existing advanced methods such as DeepLabV3+, SegNet, and UperNet. The experimental structure shows that the CNN GrabCut algorithm achieved 98% MPA and 92% MIoU, significantly surpassing the comparison model. This performance improvement is mainly due to the introduction of the CBAM module, which enhances feature extraction capability through a dual attention mechanism of channel and space. The dual path design of CBAM enables it to simultaneously capture color distribution and spatial relationship features, which is crucial for maintaining the structural integrity of the artwork. In addition, the fusion of CNN and GrabCut not only retains the precise boundary segmentation ability of GrabCut, but also compensates for the lack of computational efficiency through CNN's hierarchical feature extraction, achieving a good balance between accuracy and speed.

In terms of style transfer, the GCG system achieved a PSNR of 54dB and an SSIM of 0.92, outperforming the comparison system. A high SSIM value indicates that the generated image maintains a high degree of structural similarity with the original content while adapting to diverse artistic styles. This advantage stems from the improved GAN architecture, where residual connections effectively solve the problems of gradient vanishing and network degradation, making deeper network training possible; And semantic guided GAN achieves adaptive application of style features in semantically related regions. For example, foreground objects adopt stronger style transfer to highlight artistic expression, while the background maintains natural transitions, thereby enhancing overall visual coherence.

Despite the aforementioned advantages of the GCG

system, its computational load has slightly increased. The ablation experiment showed that the complete model requires 710MB of memory and 62ms of single frame processing time, while the baseline CNN model only requires 580MB and 45ms. This cost has resulted in significant improvements in segmentation accuracy and style transfer quality. Asynchronous parallel processing of multi-scale masks and style features through CUDA event streams further optimizes resource utilization and ensures real-time performance in high-resolution images. In the robustness test, the GCG system performed excellently under challenging conditions such as adding Gaussian noise and regional occlusion: the MIoU decrease did not exceed 4.1%, and the SSIM decrease was controlled within 0.08. This is attributed to the adaptability of GMM to noise distribution and the protective effect of residual connections on feature integrity.

In summary, the research method effectively addresses the key limitations of current art image generation systems. Innovations such as CBAM and residual connections have brought high precision and strong robustness to it, although the computational requirements have increased, it still remains within an acceptable range.

6 Conclusion

To address the challenges of handling blurred boundaries and insufficient naturalness in style in current artistic image generation systems, this study put forward an innovative system that integrated semantic segmentation and style transfer techniques. The system combined an improved CNN-GrabCut algorithm with an enhanced GAN to accurately segment object edges and merge multiple styles, thereby enabling artistic design for painting. Experimental results showed that the proposed CNN-GrabCut algorithm achieved a maximum MPA of 98% and a minimum of 96% when predicting different pixel types. The AUC value reached 0.873, the lowest F1 score was 80%, and the highest MIoU reached 92%, all outperforming the compared algorithms. The proposed GCG system demonstrated better performance in image processing, with a maximum memory usage of 710 MB and the fastest response time of 62 ms, compared to 880 MB and 81 ms of the other systems. In image quality testing, the GCG system achieved the highest PSNR of 54 dB and the highest SSIM of 92%, surpassing the maximum PSNR of 46 dB and SSIM of 84% of the compared systems. In terms of image diversity, the GCG system produced a maximum of 4000 images and 400 image categories, both higher than those of the other systems. In conclusion, the model that integrated semantic segmentation and style transfer showed strong performance in image extraction and style fusion. However, there are still limitations in the research, as it has not tested aspects such as art evaluation and aesthetic analysis. Therefore, future studies will explore the possibility of integrating aesthetic quality predictors to automatically evaluate the aesthetic quality of generated images through deep learning-based models. At the same time, a human feedback loop mechanism can be designed, where users can evaluate and provide feedback on the generated images, and the system can adjust and optimize based on this feedback.

References

[1] Wenjing X, Cai Z. Assessing the best art design based on artificial intelligence and machine learning using GTMA. Soft Computing, 2023, 27(1): 149-156. DOI: 10.1007/s00500-022-07555-1

[2] Xu J, Zhang X, Li H, Yoo C, Pan Y. Is everyone an artist? A study on user experience of AI-based painting system. Applied Sciences, 2023, 13(11): 6496-6517. DOI: 10.3390/app13116496

[3] Guo C, Lu Y, Dou Y, Wang F Y. Can ChatGPT boost artistic creation: The need of imaginative intelligence for parallel art. IEEE/CAA Journal of Automatica Sinica, 2023, 10(4): 835-838. DOI: 10.1109/JAS.2023.123555

[4] Yuan J, Zhou F, Guo Z, Li X, Yu H. HCformer: hybrid CNN-transformer for LDCT image denoising. Journal of Digital Imaging, 2023, 36(5): 2290-2305. DOI: 10.1007/s10278-023-00842-9

[5] Brophy E, Wang Z, She Q, Ward T. Generative adversarial networks in time series: A systematic literature review. ACM Computing Surveys, 2023, 55(10): 1-31. DOI: 10.1145/3559540

[6] Xie B, Li S, Li M, Liu C, Huang G, Wang G. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(7): 9004-9021. DOI: 10.48550/arXiv.2204.08808

[7] Zhou Z, Zhang J, Gong C. Hybrid semantic segmentation for tunnel lining cracks based on Swin Transformer and convolutional neural network. Computer-Aided Civil and Infrastructure Engineering, 2023, 38(17): 2491-2510. DOI: 10.1111/mice.13003

[8] Samudrala S, Mohan C K. Semantic segmentation of breast cancer images using DenseNet with proposed PSPNet. Multimedia Tools and Applications, 2024, 83(15): 46037-46063. DOI: 10.1007/s11042-023-17411-5

[9] Garg M, Ubhi J S, Aggarwal A K. Neural style transfer for image steganography and destylization with supervised image to image translation. Multimedia Tools and Applications, 2023, 82(4): 6271-6288. DOI: 10.1007/s11042-022-13596-3

[10] Li C, Taniguchi Y, Lu M, Konomi S I, Nagahara H. Cross-language font style transfer. Applied Intelligence, 2023, 53(15): 18666-18680. DOI: 10.1007/s10489-022-04375-6

[11] Chung C Y, Huang S H. Interactively transforming Chinese ink paintings into realistic images using a border enhance generative adversarial network. Multimedia tools and applications, 2023, 82(8): 11663-11696. DOI: 10.1007/s11042-022-13684-4

[12] Ugail H, Stork D G, Edwards H, Seward S C, Brooke C. Deep transfer learning for visual analysis and attribution of paintings by Raphael. Heritage Science, 2023, 11(1): 268-282. DOI: 10.1186/s40494-023-01094-0

[13] Eom T H, Lee H S. A study on the diagnosis technology for conservation status of painting cultural heritage using digital image analysis program. Heritage, 2023, 6(2): 1839-1855. DOI: 10.3390/heritage6020098

[14] Lu Y, Guo C, Dai X, Wang F Y. Generating emotion descriptions for fine art paintings via multiple painting representations. IEEE Intelligent Systems, 2023, 38(3): 31-40. DOI: 10.1109/MIS.2023.3260992

[15] Fukunaga K. Nondestructive evaluation of lined paintings by THz pulsed time-domain imaging. Heritage, 2023, 6(4): 3448-3460. DOI: 10.3390/heritage6040183

[16] Luo Z, Yang W, Yuan Y, Gou R, Li X. Semantic segmentation of agricultural images: A survey. Information Processing in Agriculture, 2024, 11(2): 172-186. DOI: 10.1016/j.inpa.2023.02.001

[17] Shi H, Dao S D, Cai J. LLMFormer: Large language model for open-vocabulary semantic segmentation. International Journal of Computer Vision, 2025, 133(2): 742-759. DOI: 10.1007/s11263-024-02171-y

[18] Chen D. Animation VR Scene Stitching Modeling Based on Genetic Algorithm. Informatica, 2024, 48(5):83-96. DOI: https://doi.org/10.31449/inf.v48i5.5364

[19] Zhang Y, Tang F, Dong W, Huang H, Ma C, Lee T Y, Xu C. A unified arbitrary style transfer framework via adaptive contrastive learning. ACM Transactions on Graphics, 2023, 42(5): 1-16. DOI: 10.1145/3605548

[20] Li H, Wang L, Liu J. Application of multi-level adaptive neural network based on optimization algorithm in image style transfer. Multimedia Tools and Applications, 2024, 83(29): 73127-73149. DOI: 10.1007/s11042-024-18451-1

[21] Meng X. Research on the Development of Modern Design through Data Mining Technology. Informatica, 2024, 48(6):59-71. DOI: 10.31449/inf.v48i6.5241

[22] Kabir A I, Mahomud L, Ald A, Ahmed R. Empowering Local Image Generation: Harnessing Stable Diffusion for Machine Learning and AI. Informatica Economica, 2024, 28(1):25-38. DOI: 10.24818/issn14531305/28.1.2024.03