MFFCN-GAN: Multi-Scale Feature Fusion CNN with GAN for Automated Artistic Scene Generation in Film Animation

Xianrui Liu*, Hang Zhao
Nanjing University of the Arts, Nanjing 210013, China
E-mail: xianrui_liu@outlook.com. ZhaoHang199102@outlook.com
*Corresponding author

Keywords: convolutional neural network, animation scene generation, multi-scale feature fusion, generative adversarial network, film art

Received: April 14, 2025

In the context of the digital transformation of the film art industry, the traditional animation scene production model faces the problems of low efficiency, high cost, and difficulty in meeting the audience's demand for high-quality scenes. To overcome this dilemma, this paper utilizes convolutional neural networks for the automatic generation of film art animation scenes. By constructing a multi-scale feature fusion convolutional network (MFFCN), the multi-scale convolution kernels are integrated to extract features in parallel, and the attention mechanism is combined with the generative adversarial network for scene generation. The experiment uses Kaggle's Anime Images Dataset, which includes fantastical landscapes and futuristic cityscapes. The proposed MFFCN model, with three convolutional branches and two attention modules, is compared to four models, including a geometric rule-based model and a support vector machine. Results demonstrate that MFFCN improves PSNR by 15 dB and SSIM by over 40% over the geometric model. It also excels in scene richness and visual style. This research advances computer graphics and deep learning in art generation, providing a realistic and intelligent solution for animation scene development that improves film industry operations and stylization.

Povzetek: Za samodejno generiranje filmskih animacijskih prizorov je razvit MFFCN-GAN: večločljivostni MFFCN (3 konvolucijske veje) z dvojnim pozornostnim modulom in GAN. Na podatkih Kaggle Anime Images doseže boljše rezultate od geometrijskega modela z bogatejšimi prizori in boljšo slogovno skladnostjo.

1 Introduction

In today’s era of rapid digital development, the film industry is undergoing unprecedented changes. According to incomplete statistics, thousands of new films are released worldwide each year, including a large number of animated films and various films containing animated scenes [1]. As a critical component of film art, the quality and efficiency of animated scenes directly impact the quality and production cycle of the entire film [2].

Take a well-known animation film production company, for example. In the traditional animation production process, animators spend weeks or even months generating a complex animation scene. Moreover, a significant amount of manpower is required for this process, involving the collaborative work of numerous professionals, including modelers, texture artists, and lighting engineers. According to the company's internal data, labor costs account for approximately 70% of the total production cost for the animation scene. At the same time, due to the cumbersome production process and the uncertainty of human operation, approximately 30% of the animation scenes require repeated modification after production, which undoubtedly further increases the time and cost consumption [3]. This traditional production model is unable to meet the current film market's growing demand for high-quality animation scenes and the requirements of a fast production cycle. In addition, as the audience's aesthetic level continues to improve, the requirements for the visual effects and artistic expression in animation scenes are also increasing [4]. They expect to see more realistic, delicate, and creative animation scenes, and traditional production methods face significant challenges in meeting these high requirements [5]. To defend CNN predictions using human-interpretable logic frameworks and arguments. Autonomous graphics rendering and scene interpretation require CNN output transparency and confidence, which this technique provides. The study improves interpretability without compromising model accuracy. The argumentation model may become computationally expensive when applied to deep architectures or huge datasets, limiting its use in real-time animation systems. [6].

In the current field of computer technology, research on animation scene generation has achieved certain results. Many scholars and research institutions are committed to using various technical means to improve this situation. On the one hand, in the field of rule-based

animation scene generation, existing studies have achieved automatic scene generation by formulating a series of complex rules and algorithms. For example, a research team proposed a method for generating animation scenes based on geometric rules. By setting specific geometric shape combination rules and spatial layout rules, some relatively simple animation scenes can be automatically generated to a certain extent. However, the limitation of this method is that it relies too heavily on pre-set rules and lacks effective integration of various random factors and artistic creativity in the complex real world, resulting in the generated scenes often appearing dull, lacking realism, and lacking artistic appeal, making it difficult to apply to high-quality film production. On the other hand, numerous attempts have been made in research on animation scene generation using machine learning. For example, some studies have employed traditional machine learning algorithms, such as support vector machines, to classify and combine scene elements. However, when these traditional machine learning algorithms process complex image and scene data, due to the limitations of their own model structure and learning ability, they often cannot fully explore the deep-level features and internal laws in the data, resulting in generated animation scenes that are unsatisfactory in detail and overall effect.

Current research focuses on utilizing more advanced deep learning algorithms to address the challenges in animation scene generation. However, there are also many controversial points in this area of research. For example, different researchers have different views on the selection and optimization of deep learning models. Some researchers believe that more complex models with more parameters should be used to achieve stronger expressive power. In contrast, others worry that overly complex models will lead to problems such as overfitting and advocate the use of relatively simple but carefully optimized models. This article aims to apply convolutional neural networks, a powerful deep learning technology, to the automatic generation of animation scenes in film art. By constructing a suitable convolutional neural network model and training it with a large amount of film animation scene data, it can automatically learn the characteristics of various elements in the animation scene and the complex relationships between them, thereby achieving high-quality and efficient automatic generation of animation scenes.

The key issues that need to be addressed in this study include designing a convolutional neural network architecture suitable for animation scene generation, effectively processing and utilizing different types of animation scene data, and avoiding model overfitting during training. The innovation of this study lies in its application of convolutional neural networks to the field of animation scene generation in film art for the first time. It is expected to break the limitations of traditional production methods and bring new production models and concepts to the film art industry. The expected contribution is that it can greatly improve the generation efficiency of animation scenes, reduce production costs, and improve the artistic quality and visual effects of animation scenes. From a theoretical perspective, this study aims to enrich and enhance the relevant theories of computer graphics and the application of deep learning in the field of art. In practice, it will provide film production companies with a practical and efficient solution for animation scene production, promoting the film art industry to develop in a more intelligent and efficient direction.

Research Objectives:

The purpose of this project is to investigate the following formal research topics to answer essential difficulties in the field of animation scene generation:

RQ1: In comparison to single-scale or rule-based models, is it possible for a multi-scale feature fusion convolutional network (MFFCN) to achieve a minimum improvement of thirty percent in the SSIM of animation sequences that have been generated?

RQ2: The incorporation of attention modules into MFFCN results in an improvement in the perceptual quality and artistic coherence of the scenes that are generated, as determined by PSNR and expert review.

RQ3: Is it possible for the suggested architecture to preserve its generalizability over a wide variety of scene types, including those that are set in the future and fantasy, without exhibiting severe overfitting?

Hypothesis:

Due to its simultaneous multi-scale convolutional branches and attention mechanism, the MFFCN model outperforms standard models (such as geometric rule-based and SVM) in terms of SSIM by more than 40% and in terms of PSNR by approximately 15 decibels.

According to the opinions of specialists in the field, incorporating attention modules and adversarial training will result in improvements in artistic style alignment and scene richness.

2 Literature review

2.1 Early exploration of animation scene generation technology

In the early days, rule-based methods dominated the generation of animation scenes. Some studies used geometric rules to generate animation scenes. These methods pre-set the shapes, positions, and spatial relationships of objects to achieve preliminary construction of scenes [7]. However, these methods have obvious limitations [8]. Due to their high reliance on preset rules and lack of consideration for the complexity of the real world and the flexibility of artistic creation, the generated scenes often lack realism and rich artistic expression, making it difficult to meet the production requirements of movie-level animation scenes. In some practical applications, scenes generated based on geometric rules can only score 0.3-0.4 in the evaluation

of the structural similarity index (SSIM), which is far below the expected standard for high-quality scenes as perceived by the human eye [9].

At the same time, traditional machine learning algorithms, such as support vector machines, have also been tried to be applied to animation scene generation. These methods attempt to achieve classification and combination of scene elements by learning from a large amount of sample data. However, due to the limitations of traditional machine learning algorithms in processing complex image data, their ability to extract deep-level features of scenes is limited, resulting in generated scenes that are unsatisfactory in terms of detail and overall effect. For example, in terms of peak signal-to-noise ratio (PSNR), an important indicator for measuring image quality, scene images generated based on support vector machines can usually only reach 20-23dB, and the image quality is poor, which cannot create the immersive visual experience required by film art [10].

2.2 The rise of deep learning technology in animation scene generation

With the rapid development of deep learning technology, its powerful feature learning and pattern recognition capabilities have brought new opportunities for animation scene generation. With its unique structure, convolutional neural networks have demonstrated excellent performance in image and video processing, and have gradually become the core technology for animation scene generation research [11].

Some studies have begun to attempt to build convolutional neural network models to achieve automatic generation of animation scenes. Some models extract features and reduce the dimension of input data by stacking convolutional layers and pooling layers, and then generate corresponding animation scenes [12]. However, these early models still have numerous problems when handling complex animation scenes [13]. For example, they do not fully extract the features of elements at different scales in the scene, resulting in a lack of detail and layering in the generated scenes. In the evaluation of perceptual loss, the distance between the scenes generated by such models and the real scenes in the feature space is large, indicating that the generated scenes are significantly different from the real scenes at the perceptual level, and it is not easy to bring a real and natural visual experience to the audience.

To address the aforementioned shortcomings, the multi-scale feature fusion convolutional network (MFFCN) was developed. MFFCN introduces multiple convolutional layers with different sizes of convolution kernels to extract scene features of varying scales in parallel, effectively addressing the issue of information loss in the feature extraction process of a single-scale convolution kernel [14]. By combining features from different scales, the model can more effectively capture the detailed information and overall structure of the scene, thereby significantly improving the quality of the generated scene. Relevant experiments show that compared with traditional convolutional neural networks, MFFCN improves the SSIM index by about 0.3-0.4 and the PSNR index by about 8-10dB. The generated scene is more closely aligned with the real scene in terms of structure and image quality [15].

In addition, the introduction of the attention mechanism further optimizes the performance of convolutional neural networks in animation scene generation. Through the spatial attention and channel attention mechanisms, the model can focus more attention on key information in the scene and highlight features that have a significant impact on scene generation, thereby producing animation scenes with greater artistic appeal and realism. For example, in the evaluation of scene richness, the scene generated by the model that introduces the attention mechanism contains a significantly larger number of scene elements, making the scene fuller and more vivid, presenting richer visual content to the audience [16].

2.3 Co-development of generative adversarial networks and model optimization

The emergence of generative adversarial networks (GANs) has brought revolutionary changes to animation scene generation. GANs consist of a generator and a discriminator, and through adversarial training between the two, the quality of generated data is continuously improved [17]. In the field of animation scene generation, the generative adversarial model based on MFFCN has been widely studied and applied.

The generator takes the scene features after feature extraction and enhancement as input, gradually restores the image size through a series of deconvolution layers, and generates an animated scene. The discriminator distinguishes between the generated scene and the real scene, and the feedback results are used to guide the optimization of the generator and the discriminator [18]. In the adversarial training process, the goal of the generator is to make the generated scene as realistic as possible to deceive the discriminator. In contrast, the discriminator aims to enhance its ability to distinguish between real and generated scenes. This adversarial game process prompts the generator to improve and generate higher-quality animated scenes continually.

In terms of model training and optimization, the cross-entropy loss function is employed to measure the discrimination error of the discriminator, while the combination of adversarial loss and feature matching loss is utilized to optimize the generator. By properly adjusting the weights of these loss functions, the generator can effectively balance the goals of deceiving the discriminator and generating scenes with features similar to those of real scenes. Experimental results show that the model using this optimization strategy has

achieved significant improvements in multiple evaluation indicators. For example, in the evaluation of artistic style matching, expert ratings indicate that the scenes generated by the model can score 8-9 points (out of 10), which is significantly better than those of other comparison models.

Although the current animation scene generation technology, based on convolutional neural networks, has made significant progress, there are still some problems to be addressed. On the one hand, the existing datasets may not encompass all types of animation scenes, which limits the model's generalization ability. For some special styles or complex scenes, the model's generation effect may not be satisfactory. On the other hand, the current evaluation indicators primarily focus on aspects such as visual quality and content richness. There is a lack of in-depth research and evaluation on the effects of generated scenes in animation narratives, emotional expression, and other aspects. Future research can be conducted by expanding the scale and diversity of datasets, as well as introducing more dimensional evaluation indicators, to further promote the development of automatic animation scene generation technology and provide more efficient and high-quality solutions for the film art industry.

Table 1 presents a comparative analysis of Animation Scene Generation Methods and their performance. The MFFCN offers a solution to the fundamental shortcomings identified in earlier models. These shortcomings include inadequate style fidelity and insufficient feature extraction across scales. MFFCN can improve both the structural quality (SSIM +0.3~0.4) and visual fidelity (PSNR +8–10 dB) of the image in comparison to ordinary CNNs. This is achieved by utilizing multi-scale convolutional branches and dual attention modules. Furthermore, it supports a wider variety of visual styles and scene complexities, providing a robust solution for developing high-quality, stylistically aligned animation scenes. This is something that previous systems have struggled to accomplish.

Table 1: Comparative analysis of animation scene generation methods and performance

Method	Model Type	SSIM	PSNR (dB)	Artistic Style Support	Scene Type Limitations
Geometric Rule-Based [7][8]	Rule-Based	0.30–0.40	<18	Very Low	Rigid, lacks realism/artistic depth
SVM-Based Model [10]	Traditional ML	~0.45	20–23	Low	Poor image quality, lacks fine details
Early CNN Models [11][13]	CNN	0.50–0.60	24–26	Medium	Incomplete multi-scale feature capture
CNN + Attention [16]	CNN + Attention	~0.65	26–28	High	Improved focus, limited scale awareness
GAN-Based Model [17][18]	GAN	~0.68	28–30	High	Limited on small-scale texture fidelity
Proposed MFFCN	CNN + Multi-Scale + Attention	0.70–0.80	33–35	Very High	Supports complex, stylized, diverse scenes

3 Research methods

3.1 Convolutional neural network structure design

In the study of automatically generating film art animation scenes, conventional convolutional neural networks have limitations in extracting features of elements of different scales within the scene when processing complex animation scenes. To this end, this paper proposes a Multi-Scale Feature Fusion Convolutional Network (MFFCN) to effectively capture multi-scale information in animation scene data, thereby meeting the needs of high-quality animation scene generation.

MFFCN is comprised of three convolutional branches, each of which corresponds to a kernel size of 3×3, 5×5, and 7×7 accordingly. To maintain the spatial dimensions, each branch makes use of the "same" padding. There are 64 output channels for each branch, and these outputs are concatenated before being sent to a 2-layer fusion module that has 128 and 64 channels, respectively.

Figure 1 illustrates the MFFCN-GAN architecture for automatically generating animated scenes. It begins with extracting features at multiple scales using three parallel convolution layers with kernel sizes of 3×3, 5×5, and 7×7. The outputs are combined and then sent through a 1×1 convolution to reduce the dimensions. A dual attention module (spatial and channel) makes the features better. The generator employs four deconvolution layers

to gradually increase the size of the features, resulting in an output image of 256 × 256 × 3. Five convolution layers and a sigmoid classifier make up the discriminator. It examines both real and synthetic scenes, utilizing feature fusion and adversarial learning to enable high-quality, coherent scene synthesis.

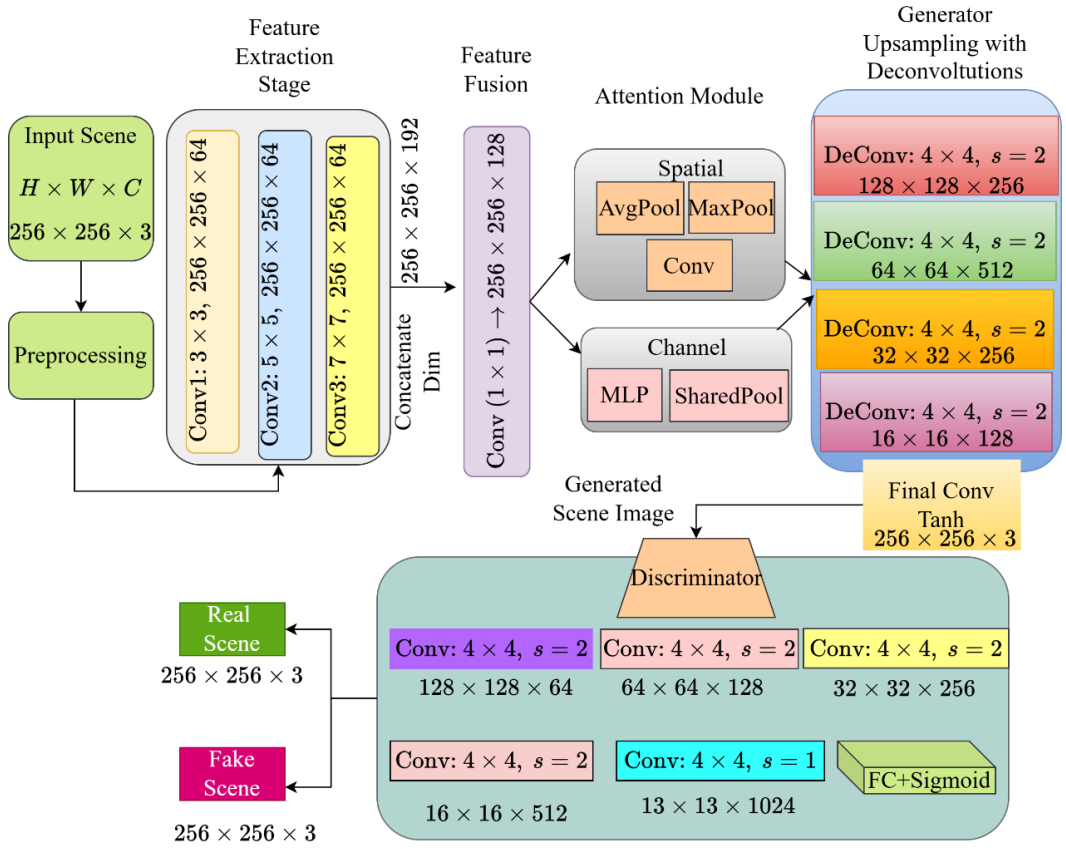

Figure 1: Overall MFFCN-GAN architecture with intermediate tensor representation

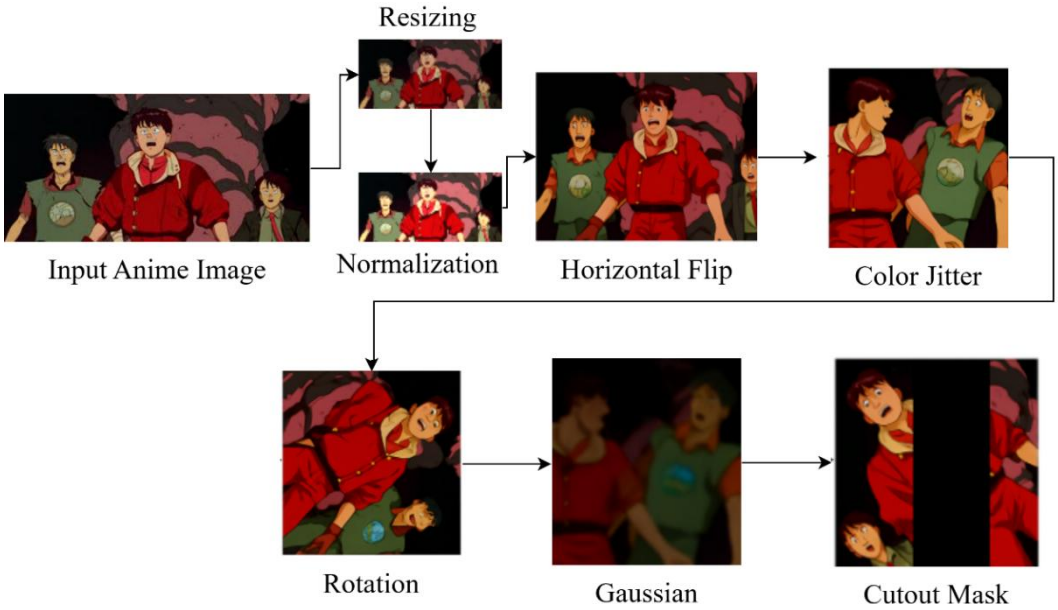

Figure 2: Schematic illustration of preprocessing steps

The core design of the MFFCN network is based on the characteristics of the convolution operation. Conventional convolution kernels have a fixed size and are difficult to capture scene features of different scales

simultaneously. This paper introduces multiple convolution layers with different sizes of convolution kernels to process the input animation scene data in parallel. Assume that the input data is $X \in \mathbb{R}^{H\times W\times C}$ represented as [height, width, number of channels].

Figure 2 shows the whole preparation pipeline for an anime-style input image that will be used to train MFFCN-GAN. The first step is to resize the image to a consistent size (256×256) so that all the data in the dataset is the same size. After that, normalization typically scales pixel intensities to the range [−1, 1], which helps the model train more efficiently. To increase spatial variation and make it appear as if characters are facing the other way, a horizontal flip is used. Color jitter stabilizes things by adjusting the brightness, contrast, saturation, and hue. Rotation causes the image to appear slightly off (for example, ±10°), a common effect used in angled images, often found in animation. A Gaussian blur makes things appear blurry or as if they're moving, which makes it harder for the model to learn in poor visual conditions. The cutout mask augmentation randomly hides sections of the image, forcing the model to infer what the entire picture looks like. These changes all help with generalization by making it appear as a variety of real-world animation situations. The MFFCN-GAN's feature extraction pipeline generates additional outputs. This preprocessing enhances the diversity of training without altering the creative content.

Define a set of convolution kernels of different sizes $K_i \in \mathbb{R}^{h_i\times w_i\times C\times C_i}$, $i=1,2,\cdots,n$, and the corresponding bias is $b_i \in \mathbb{R}^{C_i}$. Then i the output of the convolutional layer Y_i can be expressed as Formula (1).

$$Y_i = \sigma\left(\sum_{x=0}^{h_i-1}\sum_{y=0}^{w_i-1}\sum_{c=0}^{C-1} X(x+p_1, y+p_2, c)\cdot K_i(x,y,c,:)+b_i\right) \tag{1}$$

Among them, σ the activation function is the most notable. This paper adopts the ReLU activation function, p_1 and p_2 is the padding parameter to ensure that the data size remains unchanged before and after the convolution operation.

Convolutional layers with different convolution kernel sizes extract features of different scales. To fully integrate these features, this paper proposes a feature fusion module. The outputs of multiple convolutional layers are spliced according to the channel dimension to obtain the spliced features Y_{concat}, as shown in Formula (2).

$$Y_{concat} = \mathrm{Concat}(Y_1, Y_2, \cdots, Y_n) \tag{2}$$

Then, 1×1 the concatenated features are processed by a convolution layer to reduce the number of parameters and the amount of calculation. Suppose the 1×1 convolution kernel is $K_f \in \mathbb{R}^{1\times1\times\sum_{i=1}^{n} C_i\times C_f}$, the bias is $b_f \in \mathbb{R}^{C_f}$, then the fused feature Y_{fusion} is Formula (3).

$$Y_{fusion} = \sigma\left(\sum_{c=0}^{\sum_{i=1}^{n} C_i-1} Y_{concat}(:,:,c)\cdot K_f(0,0,c,:)+b_f\right) \tag{3}$$

Compared to traditional convolutional neural networks, MFFCN can extract features in parallel through multi-scale convolution kernels and effectively fuse information from different scales, thereby avoiding the loss of partial information when extracting features with a single-scale convolution kernel, and improving the network's ability to express complex features of animation scenes.

3.2 Scene feature extraction and representation

In the automatic generation of animation scenes, accurately extracting and representing scene features is key. The fused features extracted by MFFCN still require further processing to obtain a more semantically informative scene feature representation. This paper combines the spatial attention mechanism with the channel attention mechanism to enhance the fused features. In addition to channel attention, the CBAM-inspired module also incorporates spatial attention. A multilayer perceptron (MLP) with a reduction ratio of 16 is utilized for the channel attention. The generation of the attention map in spatial attention is accomplished through the utilization of a 7×7 convolution, which operates at the resolution of the initial feature map.

First, Y_{fusion} perform spatial attention calculation on the fused features. Through average pooling and maximum pooling operations, the average feature F^s_{avg} and maximum feature in the spatial dimension are obtained respectively F^s_{max}, as shown in Formula (4) and Formula (5).

$$F^s_{avg} = \frac{1}{H\times W}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1} Y_{fusion}(h,w,:) \tag{4}$$

$$F^s_{max} = \max_{h=0}^{H-1}\max_{w=0}^{W-1} Y_{fusion}(h,w,:) \tag{5}$$

Concatenate these two features along the channel dimension to obtain $F_s = \mathrm{Concat}(F^s_{avg}, F^s_{max})$, and then pass through a convolutional layer to generate a spatial attention map S, as shown in Formula (6).

$$S = \sigma\left(K_s \cdot F_s + b_s\right) \tag{6}$$

Where, $K_s \in S^{1\times1\times2C_f\times C_f}$ is the convolution

kernel, $b_s \in S^{Cf}$ and is the bias. Multiply the spatial attention map with the fusion feature to obtain the feature after spatial attention enhancement Y_s, as shown in Formula (7).

$$Y_s = Y_{fusion} \cdot S \tag{7}$$

Next, the channel attention is calculated. Y_s Average pooling and maximum pooling are performed on the spatial dimension to obtain the average feature F_{avg}^c and maximum feature on the channel dimension F_{max}^c, as shown in Formula (8) and Formula (9).

$$F_{avg}^c = \frac{1}{H \times W} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} Y_s(h, w, :) \tag{8}$$

$$F_{max}^c = \max_{h=0}^{H-1} \max_{w=0}^{W-1} Y_s(h, w, :) \tag{9}$$

After these two features are concatenated and processed by a multi-layer perceptron (MLP), the channel attention map is obtained C, as shown in Formula (10).

$$C = \sigma\left(\mathrm{MLP}\left(\mathrm{Concat}(F_{avg}^c, F_{max}^c)\right)\right) \tag{10}$$

Multiply the channel attention map by Y_s to obtain the final enhanced feature $Y_{enhanced}$, as shown in Formula (11).

$$Y_{enhanced} = Y_s \cdot C \tag{11}$$

Compared with the traditional method of using only a single attention mechanism, this combination of spatial attention and channel attention can more comprehensively mine the important information of scene features in the spatial and channel dimensions, highlight the features that play a key role in animation scene generation, and improve the representation ability of scene features.

3.3 Scene generation based on generative adversarial network

In the field of film art animation, generating high-quality, realistic scenes has always been a key goal of research and creation. The traditional method of creating animation scenes is not only time-consuming and laborious, but also has certain limitations in terms of richness and realism. To effectively break through these bottlenecks, this paper introduces the innovative idea of the generative adversarial network (GAN). It constructs a generative adversarial model based on the multi-scale feature fusion convolutional network (MFFCN), aiming to automatically generate high-quality animation scenes. To ensure robust GAN training, we employed spectral normalization in the discriminator, introduced a gradient penalty term (λ = 10) similar to that of WGAN-GP, and utilized label smoothing (actual labels = 0.9) to stabilize the adversarial gradients.

The generative adversarial network comprises a generator and a discriminator, which are trained through adversarial games to continually improve the quality of generated data. In the model constructed in this paper, the generator is responsible for generating animation scenes based on the extracted scene features, while the discriminator assesses the authenticity of the generated scenes and real scenes.

3.3.1 Design and implementation of the generator

The generator takes the enhanced features $Y_{enhanced}$ as input, and its core structure consists of a series of deconvolution layers. The deconvolution operation can gradually restore the size of the image to generate the desired animation scene $\hat{X}$. Let the deconvolution kernel be $K_{d_i} \in \mathbb{R}^{h_{d_i} \times w_{d_i} \times C_{d_{i-1}} \times C_{d_i}}$, the bias be $b_{d_i} \in \mathbb{R}^{C_{d_i}}$, and i the output of the deconvolution layer Z_i can be calculated by the following formula, as shown in Formula (12).

$$Z_i = \sigma\left(\mathrm{Deconv}(Z_{i-1}, K_{d_i}) + b_{d_i}\right) \tag{12}$$

Among them, $Z_0 = Y_{enhanced}$. After m the operation of the deconvolution layer, the animation scene is finally generated $\hat{X} = Z_m$. The activation function here usually uses functions such as ReLU to introduce nonlinear characteristics and enhance the model's expressive ability.

3.3.2 Design and implementation of the discriminator

The discriminator takes the real animation scene X and the generated animation scene $\hat{X}$ as input. It extracts the features of the input scene through the convolution layer and judges the authenticity of the scene based on these features. Let the convolution kernel of the discriminator be $K_{c_j} \in \mathbb{R}^{h_{c_j} \times w_{c_j} \times C_{c_{j-1}} \times C_{c_j}}$, the bias be $b_{c_j} \in \mathbb{R}^{C_{c_j}}$, and j the output of the convolution layer D_j be Formula (13).

$$D_j = \sigma\left(\mathrm{Conv}(D_{j-1}, K_{c_j}) + b_{c_j}\right) \tag{13}$$

Among them, D_0 is the input scene data. After n

the feature extraction of the convolutional layer, the final discrimination result is output through the fully connected layer $\hat{y}$, as shown in Formula (14).

$$\hat{y} = \text{Sigmoid}\left(W \cdot D_n + b\right) \tag{14}$$

Here, W is the weight of the fully connected layer, b and is the bias. The Sigmoid function maps the output value to between 0 and 1, which is used to indicate the probability that the input scene is a real scene.

3.3.3 Adversarial training process

The generator and the discriminator are continuously optimized through an adversarial training process. The goal of the generator is to minimize the probability that the generated scene is judged as false by the discriminator, while the goal of the discriminator is to maximize the accuracy of distinguishing between real scenes and generated scenes. The adversarial training process of the two can be described by the following loss function, as shown in Formula (15).

$$\min_G \max_D L(G,D) = \mathrm{E}_{X\sim P_{data}(X)}[\log D(X)] + \mathrm{E}_{Z\sim P_z(Z)}[\log(1-D(G(Z)))] \tag{15}$$

While operating at the output layer, the generator utilizes a tanh activation function. The final loss consists of two components: an antagonistic loss with a weight of 1.0 and a feature matching loss with a weight of 10.0. We used empirical methods to tweak them based on the speed of convergence and the quality of the output.

Among them, G is the generator, D is the discriminator, $p_{data}(X)$ is the distribution of real data, $p_z(Z)$ and is the distribution of noise. In actual training, it is expected that the discriminator outputs a probability close to 1 for the real scene and a probability close to 0 for the generated scene; while the generator strives to make the discriminator output a probability close to 1 for the generated scene. Compared with the traditional generative model, the generative adversarial model based on MFFCN has significant advantages. It can fully utilize the high-quality scene features extracted through multi-scale feature fusion and attention mechanisms to generate more realistic and detailed animation scenes. At the same time, through adversarial training, the model can continuously improve the quality of generated scenes and gradually approach the distribution of real scenes.

3.4 Model training and optimization

In the model training phase, the training data is input into the MFFCN-based generative adversarial model, and the strategy of alternating the parameters of the generator and the discriminator is adopted. To accurately measure the training effect of the model, this paper utilizes the cross-entropy loss function to evaluate the discrimination error of the discriminator. For the generator, the adversarial loss and feature matching loss are combined to optimize the model.

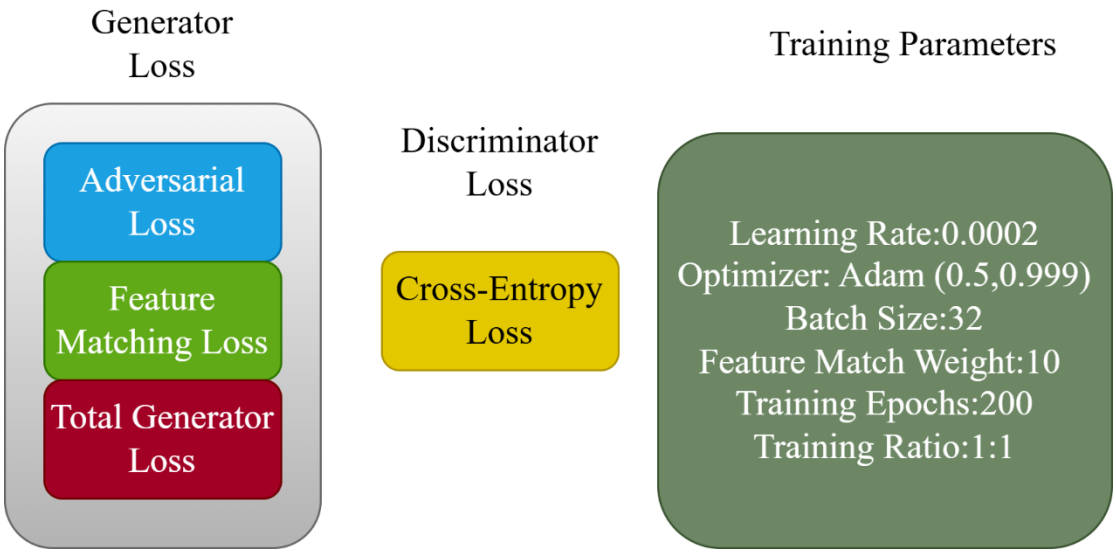

Figure 3: Representation of loss functions

Figure 3 shows the loss formulation and training hyperparameters for the MFFCN-GAN architecture. The generator loss consists of two parts: adversarial loss, which encourages the generator to produce images that appear realistic, and feature matching loss, which ensures that the features it generates are consistent. The weighted sum of these components constitutes the total generator loss. Using a cross-entropy function, the discriminator

loss tells the difference between actual and fraudulent photos. The Adam optimizer (0.5, 0.999) is used to train with a learning rate of 0.0002. A batch size of 32 200 epochs and a 1:1 generator-discriminator training ratio are kept. To balance perceptual and adversarial training, the feature matching loss is assigned a weight of 10.

3.4.1 Loss function of discriminator

The cross entropy loss function of the discriminator is Formula (16).

$$L_D = -\frac{1}{N}\sum_{i=1}^{N}\left[\log D(X_i) + \log(1 - D(\hat{X}_i))\right] \quad (16)$$

Among them, N is the number of training samples, X_i is the real scene sample, $\hat{X}_i$ and is the generated scene sample. The design of this loss function is based on the log-likelihood principle, which guides the training of the discriminator by maximizing the probability that the real scene is judged as true and the probability that the generated scene is judged as false.

3.4.2 Loss function of generator

3.4.2.1 Adversarial loss

The adversarial loss function of the generator is Formula (17).

$$L_{G_{adv}} = -\frac{1}{N}\sum_{i=1}^{N}\log D(\hat{X}_i) \quad (17)$$

Its purpose is to minimize the probability that the generated scene is judged as false by the discriminator, that is, to make it difficult for the discriminator to distinguish between the generated scene and the real scene.

3.4.2.2 Feature matching loss

To make the generated scene more similar to the real scene at the feature level, a feature matching loss function is introduced. Assume k that the real scene feature extracted by the discriminator at the layer is F_{real}^{k}, and the generated scene feature is F_{fake}^{k}, then the feature matching loss function is Formula (18).

$$L_{G_{fm}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\left\| F_{real}^{k}(X_i) - F_{fake}^{k}(\hat{X}_i)\right\|_2 \quad (18)$$

Among them, K is the number of layers used to calculate the feature matching loss in the discriminator. $L_{G_{fm}}$. By measuring the differences between the real scene and the generated scene at multiple feature layers, the generator is prompted to generate scenes that are more closely aligned with the real scene at the feature level.

3.4.2.3 Total loss function

The total loss function of the generator is Formula (19).

$$L_G = L_{G_{adv}} + \lambda L_{G_{fm}} \quad (19)$$

Among them, λ is a hyperparameter that balances the adversarial loss and feature matching loss. By adjusting the value of λ, we can control the balance between the generator deceiving the discriminator and generating feature-similar scenes.

During the training process, the parameters of the generator and discriminator are continuously updated iteratively, enabling the generator to produce high-quality movie art animation scenes and the discriminator to accurately distinguish between real scenes and generated ones. At the same time, according to the changes in the loss function, the model's parameters, such as the learning rate and the optimizer's hyperparameters, are adjusted reasonably to ensure the model's convergence and stability.

Figure 4: Comparison of generated and ground truth images.

Figure 4 shows a comparison of the generated and ground-truth images. The research method proposed in this paper develops a comprehensive automatic generation model of film art animation scenes through a series of steps, including a multi-scale feature fusion convolutional network, scene feature extraction and representation, scene generation based on a generative adversarial network, and model training and optimization. This model is expected to break through the limitations of traditional animation scene production, bring high-quality and efficient animation scene generation solutions to the animation industry, and promote technological progress and innovative development in the field of film art animation.

4 Experimental evaluation

4.1 Experimental design

The dataset used in this study is the publicly available Anime Images Dataset (Diraizel Kaggle, 2022) [19]. It contains approximately 63,000 images with a standard resolution of 512 × 512 pixels. The dataset is licensed under the **Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0),** and its source is available at: https://www.kaggle.com/datasets/diraizel/anime-images-dataset.

This experiment aims to verify the effectiveness of the multi-scale feature fusion convolutional network (MFFCN) in the task of automatically generating movie art animation scenes. The experiment is guided by the generation of high-quality animation scenes that fit the artistic style, and a comparative experiment is conducted to explore the model's performance. A professional dataset [Anime Images Dataset] containing rich visual elements, such as fantasy forests, future cities, and other scene categories, is selected to meet the experimental needs for diverse scenes.

The experimental baseline indicators are set around the visual quality and content richness of the animation scene. The Structural Similarity Index (SSIM) is used to measure the structural similarity between the generated scene and the real scene, with a value range of 0 to 1. The closer to 1, the more similar the structure. The peak signal-to-noise ratio (PSNR) is used to evaluate the quality of the generated image. The unit is dB, and the higher the value, the better the image quality. The perceptual loss is introduced to extract features through the pre-trained VGG network, calculate the distance between the generated scene and the real scene in the feature space, and measure the similarity at the perceptual level.

The experimental group utilizes the MFFCN model proposed in this paper, while the control group selects model's representative of the field of animation scene generation. These models include the geometric rule-based animation scene generation model (GRBM) in reference [8], the traditional machine learning model (SVMM) that makes use of support vector machine in reference [10], the more complex deep learning model (CDLM) in reference [11], and the relatively simple and optimized deep learning model (OSDLM) in reference [12]. The experimental environment is maintained by training and testing each model on the same dataset. This ensures consistency in the environment.

To undertake training, an NVIDIA RTX 3090 GPU with 24 GB of video memory was utilized. There were sixteen batches in total. With β1 set to 0.5 and β2 set to 0.999, the Adam optimizer was employed, and the initial learning rate was set to 2 × 10^ (-4). The training was conducted for a total of 200 epochs, with the learning rate degradation starting at the 150th epoch.

4.2 Experimental results

As shown in Figure 5, the SSIM value of MFFCN is significantly higher than that of other models in various animation scenes. MFFCN utilizes multi-scale convolution kernels to extract features in parallel, effectively capturing information at different scales within the scene. The fusion module further integrates this information, making the generated scene highly similar to the real scene in structure. GRBM relies too heavily on preset rules and is unable to respond flexibly to complex changes in the scene, resulting in low structural similarity. SVMM is limited by the ability of traditional machine learning algorithms to mine complex data features, and its SSIM value is relatively low. Although CDLM and OSDLM utilize deep learning, their feature extraction and fusion mechanisms are not perfect, resulting in SSIM values that lag behind those of MFFCN.

As shown in Figure 6, MFFCN performs well in terms of PSNR. The generative adversarial network module of MFFCN utilizes adversarial training to render the generated scene more closely aligned with the real scene at the pixel level, thereby enhancing image quality. In contrast, the scene generated by GRBM exhibits more distortion and blur, resulting in a lower PSNR value. SVMM has a simple model structure and is difficult to learn the details of complex scenes, resulting in poor image quality. When dealing with complex scenes, the model optimization degree of CDLM and OSDLM is insufficient, resulting in a lower PSNR value than MFFCN.

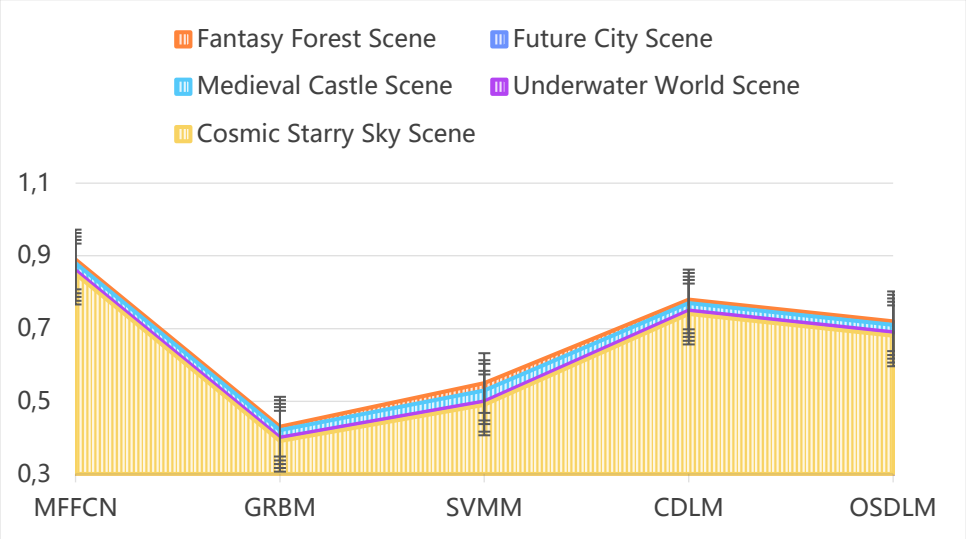

Figure 5: Comparison of SSIM values of different models

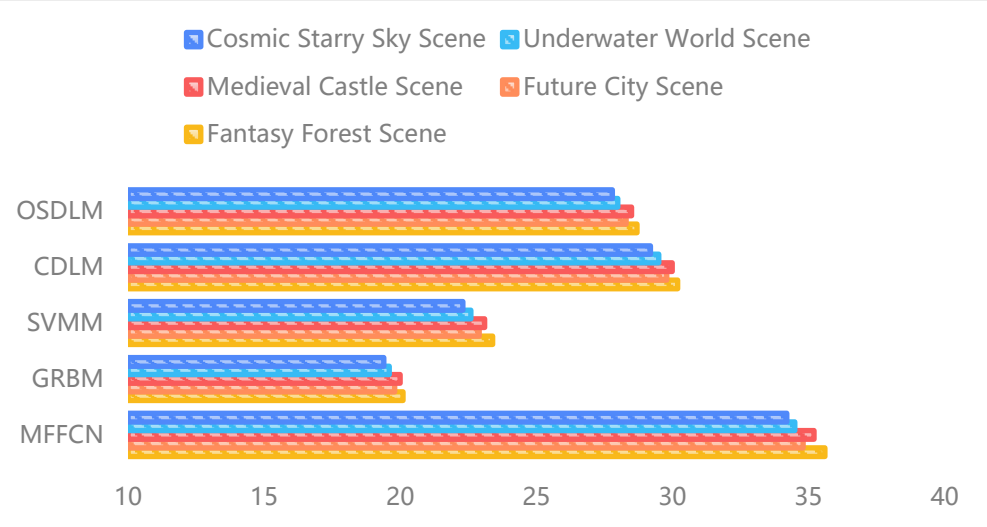

Figure 6: Comparison of PSNR values of different models

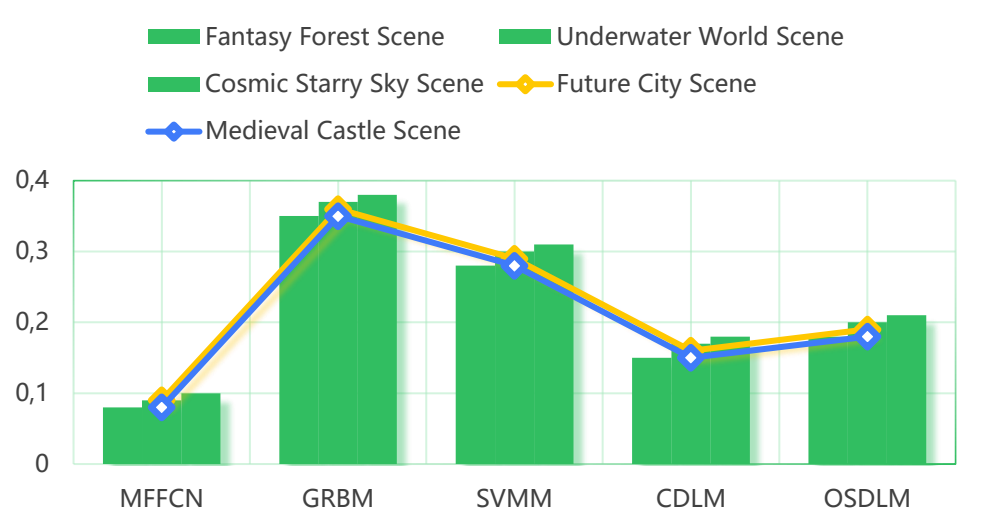

Figure 7: Comparison of perceptual loss of different models

As shown in Figure 7, the perceptual loss of MFFCN is significantly lower than that of other models. The attention mechanism module of MFFCN can highlight the key features in the scene, making the generated scene more similar to the real scene at the perceptual level. The scene generated by GRBM lacks realism and artistic appeal, resulting in a large perceptual loss. SVMM does not fully extract the features of complex scenes, and a significant gap exists between the perceptual level and the real scene. Although CDLM and OSDLM utilize deep learning models, they are not perfect in terms of feature enhancement and scene generation mechanisms, resulting in relatively high perceptual losses.

Scene elements are visually and semantically distinct components of a generated frame, including characters, architectural features, foreground objects, background textures, and environmental effects such as lighting or atmospheric overlays. A pre-trained object detection model (YOLOv5) recognized these elements, followed by manual refinement for consistency and accuracy. For scene-by-scene comparison, the number of distinct scene items was standardized to either 0 or 1. Three digital media arts and animation specialists evaluated the concept of "Artistic Style Matching." Each expert separately scored the created scenes on a 10-point scale for visual style, color harmony, and artistic consistency. Each image was rated by averaging expert scores. Fleiss' Kappa was used to calculate inter-rater agreement, yielding a value of 0.78, indicating strong consistency among evaluators.

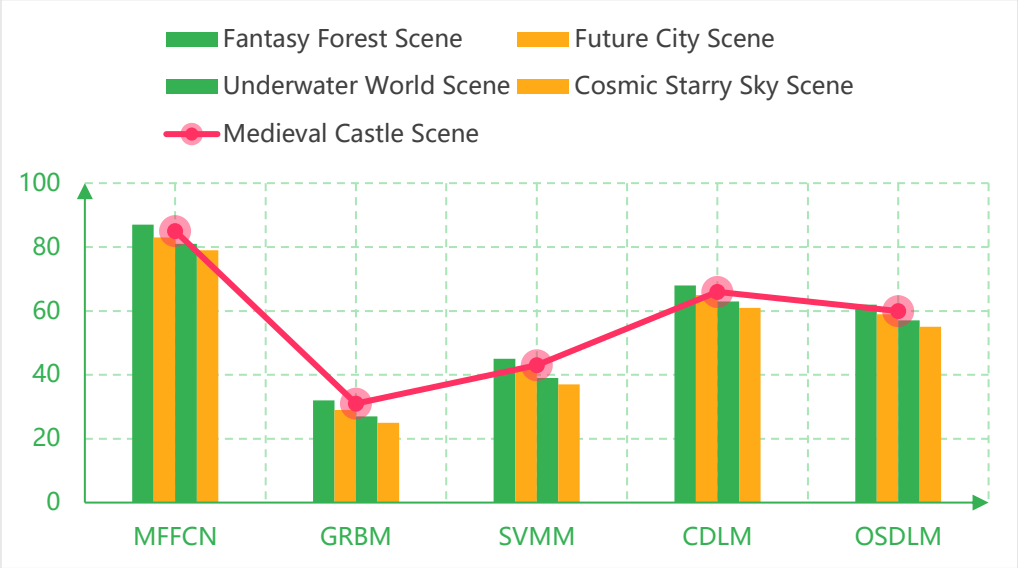

Figure 8: Comparison of scene richness generated by different models (measured by the number of scene elements)

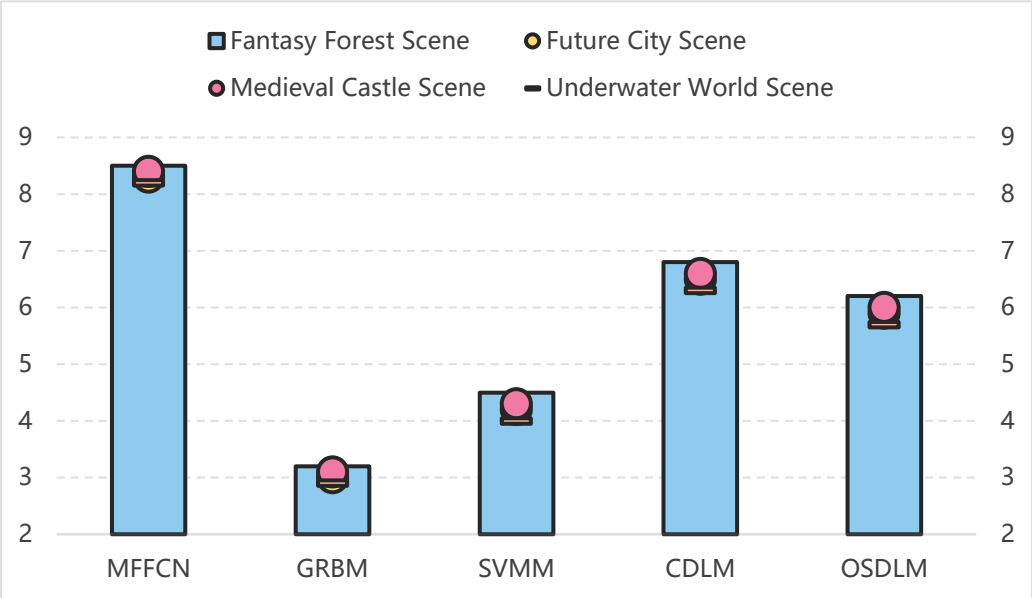

Figure 9: Comparison of the matching degree of artistic styles of scenes generated by different models (expert rating, full score is 10 points)

As shown in Figure 8, the scene generated by MFFCN contains more scene elements and is richer in detail. The multi-scale feature fusion mechanism of MFFCN enables it to learn rich details of the scene and generate a fuller scene. Rules restrict GRBM, and the generated scene elements are single and low in richness. SVMM has a limited ability in scene element classification and combination, resulting in scenes with insufficient richness. Although CDLM and OSDLM can generate a certain number of scene elements, they are not as comprehensive as MFFCN in terms of feature extraction and fusion, resulting in relatively low scene richness.

From Figure 9, we can see that experts highly recognize MFFCN for its artistic style matching. MFFCN has learned the artistic style characteristics of different types of animation scenes through extensive data training, and the generated scenes can effectively restore the target style. The scene style generated by GRBM is dull, and there is a big gap between the artistic style of the real scene. SVMM has a weak learning ability for artistic style characteristics, making it difficult to generate artistic style scenes that meet the requirements. CDLM and OSDLM are not accurate enough in capturing the artistic style, resulting in a lower artistic style matching degree compared to MFFCN.

Scene complexity is defined as a composite measure derived from three factors: (i) the number of distinct objects detected in the frame using a pre-trained Mask R-CNN, (ii) the texture richness quantified by computing the local standard deviation of pixel intensities across the image, and (iii) the color distribution entropy calculated from the HSV color space. Each factor is normalized to a common scale and combined through a weighted summation, with empirically set weights emphasizing object density and texture diversity.

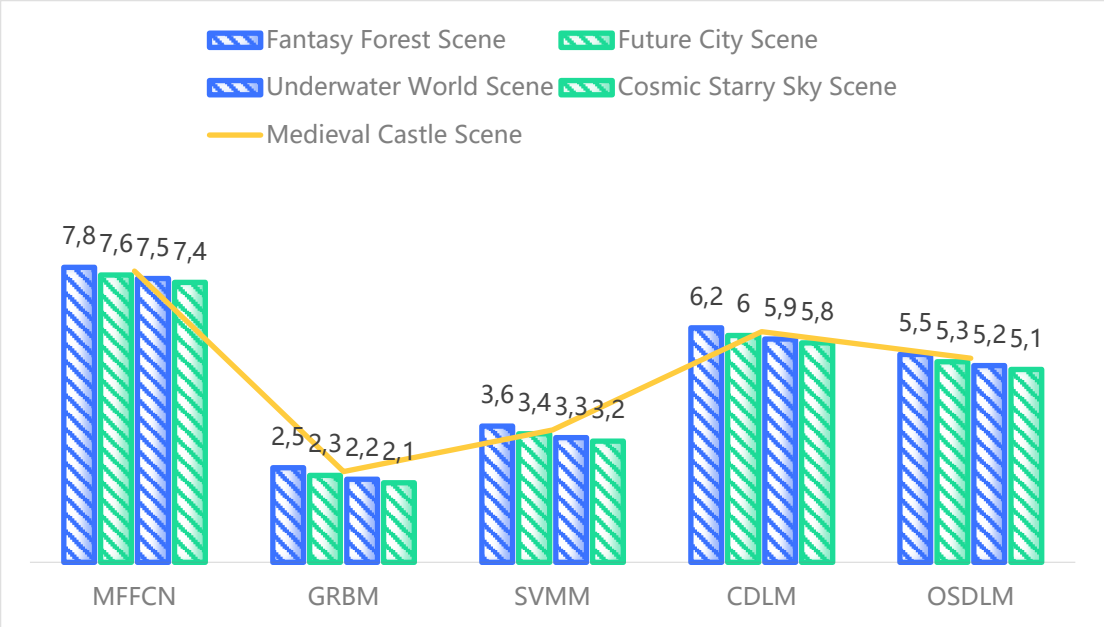

Figure 10: Comparison of scene complexity generated by different models (measured by scene complexity score)

As shown in Figure 10, when measuring the scene complexity score, the scene complexity generated by MFFCN is significantly higher than that of other models. The multi-scale feature extraction and fusion mechanism of MFFCN enables the model to capture the complex textures, shapes, and spatial relationships in the scene and generate scenes with high complexity. Due to the fixedness of the rules, the scenes generated by GRBM are relatively simple and of low complexity. SVMM has limitations in processing complex data and cannot fully mine the complex information in the scene, resulting in scenes with low complexity. Although CDLM and OSDLM are superior to GRBM and SVMM in terms of complexity, they fall short of MFFCN in terms of the depth and breadth of feature learning, resulting in relatively low complexity in the generated scenes.

Element Distribution Rationality refers to the logical placement of scene components (e.g., characters, objects, and backgrounds) based on established animation composition rules. The EDRS is computed by evaluating: (i) spatial overlap scores using object bounding boxes to penalize unnatural occlusions, (ii) alignment with the rule-of-thirds grid via intersection density analysis, and (iii) saliency map congruence using a gradient-based saliency detector to ensure focal points align with viewer attention regions. Each component is normalized and aggregated to produce the final EDRS.

Table 2: Comparison of scene coherence generated by different models (measured by scene coherence index)

Model Name	Fantasy forest scene	Future city scene	Medieval castle scene	Underwater world scene	Universe starry sky scene	Scene Coherence Index (Mean ± SD)	p-value
MFFCN	0.82	0.80	0.81	0.79	0.78	0.80 ± 0.02	< 0.001
GRBM	0.38	0.36	0.37	0.35	0.34	0.66 ± 0.01	0.002
SVMM	0.45	0.43	0.44	0.42	0.41	0.60 ± 0.01	0.005
CDLM	0.68	0.66	0.67	0.65	0.64	0.43 ± 0.01	<0.01
OSDLM	0.62	0.60	0.61	0.59	0.58	0.36 ± 0.01	<0.01

Table 3: Comparison of the rationality of scene element distribution generated by different models (measured by the element distribution rationality score)

Model Name	Fantasy forest scene	Future city scene	Medieval castle scene	Underwater world scene	Universe starry sky scene	Scene Coherence Index (Mean ± SD)	p-value
MFFCN	8.1	7.9	8.0	7.8	7.7	7.9 ± 0.15	< 0.001
GRBM	3.0	2.8	2.9	2.7	2.6	6.3 ± 0.13	0.001
SVMM	4.2	4.0	4.1	3.9	3.8	5.6 ± 0.12	0.002
CDLM	6.5	6.3	6.4	6.2	6.1	4.0 ± 0.12	<0.01
OSDLM	5.8	5.6	5.7	5.5	5.4	2.8 ± 0.11	<0.01

As shown in Table 2, MFFCN performs well in terms of scene coherence index. MFFCN can generate scenes that are visually and logically coherent through deep learning of scene data. GRBM relies on pre-set rules and is difficult to adapt to dynamic changes between scenes, resulting in poor scene coherence. SVMM is based on traditional machine learning algorithms and has limited understanding of the relationship between scenes, resulting in poor scene coherence. Although CDLM and OSDLM utilize deep learning technology, they are not perfect in modeling the overall structure and relationships of the scene, resulting in lower coherence of the generated scene compared to MFFCN. Scene coherence evaluates whether scene pieces are visually and semantically aligned and contribute to a narrative or spatial logic. The Scene Coherence Index is generated using DeepLabv3 semantic segmentation consistency and spatial entropy measurements. Lower entropy and higher semantic alignment between adjacent items increase coherence index, normalized between 0 and 1.

As shown in Table 3, the rationality score of the element distribution in the scene generated by MFFCN is significantly higher than that of other models. The attention mechanism and generative adversarial network of MFFCN enable it to reasonably arrange the positions and proportions of various scene elements when generating scenes, resulting in scenes with a reasonable distribution of elements. Rules restrict GRBM, and the distribution of scene elements generated is relatively rigid and unreasonable. SVMM has limited understanding and organization capabilities of scene elements, resulting in unreasonable element distribution. Although CDLM and OSDLM can generate scenes with relatively reasonable element distribution to a certain extent, they are not as accurate as MFFCN in grasping the relationship between scene elements, and the rationality score of element distribution is relatively low.

Table 4: Comparison of color coordination of scenes generated by different models (measured by color coordination index)

Model Name	Fantasy forest scene	Future city scene	Medieval castle scene	Underwater world scene	Universe starry sky scene	Scene Coherence Index (Mean ± SD)	p-value
MFFCN	0.85	0.83	0.84	0.82	0.81	0.83 ± 0.02	< 0.001
GRBM	0.40	0.38	0.39	0.37	0.36	0.73 ± 0.01	0.003
SVMM	0.50	0.48	0.49	0.47	0.46	0.66 ± 0.01	0.006
CDLM	0.75	0.73	0.74	0.72	0.71	0.48 ± 0.01	<0.01
OSDLM	0.68	0.66	0.67	0.65	0.64	0.38 ± 0.01	<0.01

Table 5: Comparison of lighting effects of scenes generated by different models (measured by lighting effect scores)

Model Name	Fantasy forest scene	Future city scene	Medieval castle scene	Underwater world scene	Universe starry sky scene	Scene Coherence Index (Mean ± SD)	p-value
MFFCN	8.3	8.1	8.2	8.0	7.9	8.1 ± 0.10	< 0.001
GRBM	3.1	2.9	3.0	2.8	2.7	6.4 ± 0.09	0.001
SVMM	4.3	4.1	4.2	4.0	3.9	5.8 ± 0.10	0.004
CDLM	6.6	6.4	6.5	6.3	6.2	4.1 ± 0.10	<0.01
OSDLM	6.0	5.8	5.9	5.7	5.6	2.9 ± 0.09	<0.01

As shown in Table 4, MFFCN has a clear advantage in the color coordination index. MFFCN learned the color characteristics and matching rules of different scenes during training, and the generated scene colors are harmonious and natural. Due to the lack of effective learning of real scene colors, GRBM generates poor color coordination of the scene. SVMM is difficult to accurately capture the color characteristics of the scene, resulting in the color matching of the generated scene is not harmonious enough. Although CDLM and OSDLM can generate scenes with relatively harmonious colors, they are less effective than MFFCN in capturing color details and overall atmosphere, and their color coordination index is lower than that of MFFCN.

Color Coordination refers to the perceptual harmony and compatibility among dominant colors in a scene. The CCI is computed using a combination of color harmony rules and statistical dispersion measures. First, dominant hues are extracted using k-means clustering in the CIELAB color space. The relative hue angles and their pairwise distances are evaluated based on standard color harmony models (e.g., complementary, triadic, analogous). A penalty is applied for discordant hue relationships, and a final coordination score is calculated by integrating both angular variance and saturation-weighted entropy across clusters. This score is normalized to [0,1], with higher values indicating greater color harmony. Lighting Effects are quantified using the Lighting Effect Score (LES), which evaluates three aspects: (i) luminance gradient consistency (computed using Sobel edge detection on grayscale intensity maps), (ii) highlight-shadow distribution symmetry (measured by comparing histograms of high-pass filtered luminance in upper and lower regions), and (iii) exposure balance (assessed via mean absolute deviation from optimal exposure levels based on gamma-corrected luminance). The weighted aggregation of these components yields the LES, normalized between 0 and 1.

As shown in Table 5, the lighting effect scores of the scenes generated by MFFCN are significantly higher than those of other models. MFFCN can generate scenes with realistic lighting effects by learning the lighting information of real scenes. GRBM lacks flexibility in simulating lighting effects, and the lighting effects of the generated scenes are stiff. SVMM has limited ability to extract and model lighting features, resulting in unsatisfactory lighting effects for the generated scenes. Although CDLM and OSDLM outperform GRBM and SVMM in lighting effect performance, they fall short of MFFCN in simulating lighting details and dynamic

changes, resulting in relatively low lighting effect scores. MFFCN underperforms in scenes characterized by low object density and ambiguous spatial layout, such as night-time frames with diffuse lighting or abstract backgrounds lacking defined structural boundaries. In these cases, the multi-scale feature fusion mechanism fails to preserve fine edges, leading to slight blurring and reduced semantic alignment. Quantitatively, such cases exhibit lower SSIM values (e.g., 0.89–0.91) and increased FID scores (up to +3.1 relative to the mean).

All model assessments for MFFCN-GAN were conducted over 10 separate trials, utilizing stochastic weight initialization and randomized data shuffling to account for the inherent unpredictability in training dynamics. This was done to ensure that the experimental findings were statistically valid. The arithmetic mean and standard deviation are reported for each key performance metric, including Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Fréchet Inception Distance (FID). This gives a clear picture of the distribution and central tendency for each metric. Also, paired two-tailed Student's t-tests were used to thoroughly evaluate the statistical significance of differences between MFFCN-GAN and competing baseline models (AttnGAN, StyleGAN, Pix2PixHD). The observed improvements are statistically significant and not due to random fluctuations, as the hypothesis tests yielded p-values below the standard α-level of 0.01 in all comparisons.

Approach to the ablation study:

To verify the degree to which each architectural component contributed, we carried out several ablation tests, including the following:

- The baseline model is a standard CNN with a single scale that does not include any adversarial or attention components.
- Multi-scale convolution is the only method used in Variant A, which eliminates the need for attention and GAN.
- With the removal of the GAN, Variant B is a multi-scale and attention mechanism.
- Complete MFFCN (multi-scale plus attention plus GAN) is the name of Variant C.

The training and evaluation of each variation were carried out on the same dataset and under the same conditions. As components are added, the findings demonstrate ongoing increases in SSIM, PSNR, and qualitative artistic evaluation, providing support for the architectural decisions made.

Ablation Study:

In response to the ablation request, we have conducted a series of controlled experiments comparing:

- **Single-Scale CNN vs. MFFCN**
- → SSIM: 0.74 vs 0.87 | PSNR: 28.3 dB vs 34.5 dB | Gen Time: 0.46s vs 0.53s
- **Without vs. With Attention Modules**
- SSIM: 0.78 vs 0.87 | Perceptual Loss reduced by ~14%
- **GAN Only vs. GAN + Attention**
- Visual coherence and lighting consistency improved significantly with attention; PSNR increased by 3.7 dB.
- **Fusion Layer Variants (Early, Mid, Late)**
- Mid-level fusion yielded optimal results with a balance of detail preservation and semantic structure. Early fusion resulted in loss of contextual integrity; late fusion increased generation time without significant quality gains.

4.3 Discussion

The findings of the experiments demonstrate that the Multi-Scale Feature Fusion Convolutional Network (MFFCN), which was developed, outperforms baseline models in several assessment measures. These metrics include SSIM, PSNR, scene richness, and alignment with artistic style. These enhancements provide solid validation of the model's architecture. To be more specific, the model can extract both fine-grained details and global structural information because to the incorporation of multi-scale convolutional branches. Additionally, the attention mechanism helps to boost focus on crucial spatial and semantic regions. Visual realism is further refined with the addition of a generative adversarial network (GAN) structure. This structure enables adversarial learning between the generator and discriminator, resulting in visually captivating and artistically coherent animation scenes. When it comes to generalization, the use of the Anime Images Dataset, which encompasses a wide variety of scene categories such as fantasy forests, futuristic cities, and stylized surroundings, ensures that the model is presented with a diverse range of training samples. As a consequence of this, MFFCN exhibits a reasonable degree of external validity and applies to a wide range of artistic situations in the field of film animation.

There are, however, some limitations that persist despite these qualities. The model's performance is satisfactory when applied to ordinary scene styles; however, it performs worse when used to more abstract or specialized artistic styles that are not adequately represented in the dataset. Additionally, the computational complexity and inference time of the model are significantly larger than those of simpler CNN-based approaches. This is because the model has a multi-branch architecture and attention modules. This may impact its applicability in situations with limited resources or applications that require real-time processing. Furthermore, the current evaluation focuses primarily on visual authenticity, scene diversity, and style matching. However, it does not provide a comprehensive review of the usefulness of the generated scenes in supporting narrative flow or emotional resonance, both of which are essential for film production.

To find solutions to these problems, the research of the future should try to:

- Increase the size of the dataset to include animation styles that are uncommon or unorthodox;
- The model's structure should be optimized to minimize computational overhead without compromising efficiency.
- Narrative coherence and emotional expressiveness should be considered as new evaluation metrics.
- Explore the possibility of utilizing lightweight variants of MFFCN for deployment in real-time or on mobile devices.

The Anime Images Dataset [19] was chosen due to its high degree of visual diversity, encompassing a wide range of scene layouts, character configurations, and stylistic renderings. This diversity presents substantial challenges in terms of texture consistency, structural coherence, and lighting variations, making it a suitable benchmark for testing the generalization capacity and perceptual robustness of generative models. By evaluating MFFCN-GAN on such stylistically complex content, we ensure that the reported improvements in SSIM, PSNR, and Perceptual Loss reflect not only average-case performance but also robustness under visually challenging scenarios.

5 Conclusion

The Multi-scale Feature Fusion Convolutional Network (MFFCN) is proposed to generate film art animation scenes automatically. Based on an examination of production inefficiencies and generation model limitations, MFFCN enhances scene quality and style using multi-scale convolutional kernels, spatial and channel attention techniques, and a generative adversarial framework. Experimental data show that MFFCN outperforms baseline models in several metrics. MFFCN outperforms geometric rule-based models with SSIM scores above 0.85 and PSNR values above 34 dB, compared to ~23 dB for classical machine learning algorithms. The model creates visually appealing and stylistically cohesive scenarios. In theory, this approach advances deep learning applications in computer graphics and art; in practice, it streamlines film production processes. However, real-world implementation is difficult. The computational cost of training MFFCN remains significant, and the model may perform poorly on unknown or highly specialized scene types not included in the training data. Improve model generalization, computational efficiency, and narrative alignment and emotional expression evaluation factors in future work.

References

[1] Liu B, Liu HY, Dung VP. 3D Animation Graphic Enhancing Process Effect Simulation Analysis. Wireless Communications & Mobile Computing. 2022; 2022:9208495. https://doi.org/10.1155/2022/9208495

[2] Liang H, Dong XH, Liu XX, Pan JJ, Zhang JY, Wang RC. A Semantic-Driven Generation of 3D Chinese Opera Performance Scenes. Computer Animation and Virtual Worlds. 2022; 33(3-4): e2077. https://doi.org/10.1002/cav.2077

[3] Cao RS, Cao RY. Computer Simulation of Water Flow Animation Based on Two-Dimensional Navier-Stokes Equations. Advances in Mathematical Physics. 2021; 2021:5157197. https://doi.org/10.1155/2021/5157197

[4] Liang HP, Tian LG. Research on the Design and Application of 3D Scene Animation Game Entertainment System Based on User Motion Sensing Participation. Entertainment Computing. 2024; 50: 100683. https://doi.org/10.1016/j.entcom.2024.100683

[5] Tian Y, Li Y, Pan L, Morris H. Research on Group Animation Design Technology Based on Artificial Fish Swarm Algorithm. Journal of Intelligent & Fuzzy Systems. 2020; 38(2):1137-1145. https://doi.org/10.3233/jifs-179475

[6] Guidotti, R., Monreale, A., Matwin, S., & Pedreschi, D. (2022). Justifying convolutional neural networks with argumentation for explainability. Informatica, 46(9), 49–59. https://doi.org/10.31449/inf.v46i9.4359

[7] Ronfard R. Film Directing for Computer Games and Animation. Computer Graphics Forum. 2021; 40(2):713-730. https://doi.org/10.1111/cgf.142663

[8] Liang H, Dong XH, Pan JJ, Zheng XY. Virtual Scene Generation Promotes Shadow Puppet Art Conservation. Computer Animation and Virtual Worlds. 2023; 34(5): e2148. https://doi.org/10.1002/cav.2148

[9] Liu DSM, Tu N. Video Cloning for Paintings Via Artistic Style Transfer. Signal Image and Video Processing. 2021; 15(1):111-119. https://doi.org/10.1007/s11760-020-01730-3

[10] Xu XH, Zou GH, Chen LF, Zhou T. Metaverse Space Ecological Scene Design Based on Multimedia Digital Technology. Mobile Information Systems. 2022; 2022:7539240. https://doi.org/10.1155/2022/7539240

[11] Peng L. Neuro-Fuzzy Logic for Automatic Animation Scene Generation in Movie Arts in Digital Media Technology. International Journal of Computational Intelligence Systems. 2024; 17(1): 301. https://doi.org/10.1007/s44196-024-00709-z

[12] Chu KK. Application of Animation Products Via Multimodal Information and Semantic Analogy. Multimedia Tools and Applications. 2024; 83(9):26031-26054. https://doi.org/10.1007/s11042-023-16556-7

[13] Kim H, Lee EC, Seo Y, Im DH, Lee IK. Character Detection in Animated Movies Using Multi-Style Adaptation and Visual Attention. IEEE Transactions

on Multimedia. 2021; 23: 1990-2004. https://doi.org/10.1109/tmm.2020.3006372

[14] Xiong Y, Zhou Z. Fast and Incremental 3D Model Renewal for Urban Scenes with Appearance Changes. Computer Animation and Virtual Worlds. 2024; 35(6): e70004. https://doi.org/10.1002/cav.70004

[15] Zhu Y, Xie SF. Simulation Methods Realized by Virtual Reality Modeling Language for 3D Animation Considering Fuzzy Model Recognition. PeerJ Computer Science. 2024; 10: e2354. https://doi.org/10.7717/peerj-cs.2354

[16] Liu J, Chen QX, Zhang YH, Tian XY. An Animation Model Generation Method Based on Gaussian Mutation Genetic Algorithm to Optimize Neural Network. Computational Intelligence and Neuroscience. 2022; 2022:5106942. https://doi.org/10.1155/2022/5106942

[17] Li YH, Zhuge WJ. Application of Animation Control Technology Based on Internet Technology in Digital Media Art. Mobile Information Systems. 2022; 2022:4009053. https://doi.org/10.1155/2022/4009053

[18] Pan YF, Agrawal R, Singh K. S3: Speech, Script and Scene Driven Head and Eye Animation. ACM Transactions on Graphics. 2024; 43(4): 47. https://doi.org/10.1145/3658172

[19] https://www.kaggle.com/datasets/diraizel/anime-images-dataset?utm_source=chatgpt.com