International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 1440-1448

https://doi.org/10.70517/ijhsa464130

AIGC and Graph Neural Network Fusion for Intelligent Generation and Dynamic Processing of Animated Scenes

Lei Yang[1,*] and Zhenyao Jin[1]
[1] School of Design and Art, Shandong Huayu University of Technology, Dezhou, Shandong, 253034, China
Corresponding authors: (e-mail: 13793450987@163.com).

Abstract Traditional animation scene generation methods rely on manual design, which is inefficient and difficult to meet the rapid demand for high-quality content in the modern animation industry. The rise of artificial intelligence content generation technology brings new opportunities for animation production, and graph neural networks show strong advantages in processing complex relational data. In this paper, we propose an animation complex scene generation and dynamic processing method that combines AIGC and graph neural networks. The method constructs a scene graph generation model based on two-layer graph convolutional network, divides the scene objects into primary and secondary objects through the object layering module, and fuses the visual, spatial and semantic features using the two-layer feature extraction module to realize the intelligent generation and optimized processing of complex animation scenes. The study is validated on a self-built ASG dataset, which contains 52k labeled animated scene images, covering 154 object classes and 70 relation classes. The experimental results show that the proposed method achieves an accuracy rate of 95.94% in the scene graph recognition task, which is 7.08 percentage points higher than that of the recurrent consistent generation adversarial network method; in terms of the animated scene generation rate, the rate of generating 100 images is 13.78 min per image, which is significantly better than that of the comparison method; in the evaluation of the effectiveness of the deblurring process, the PSNR value on the validation set reaches 35.44 dB, and the SSIM value is 0.923. The study shows that the method effectively improves the generation quality and processing efficiency of complex scenes in animation production, and provides technical support for the intelligent development of animation industry.

Index Terms AIGC, graph neural network, animation scene generation, two-layer graph convolutional network, complex scene, dynamic processing

I. Introduction

Animation production is an art form that creates a visual effect of coherent motion by presenting a series of still images in succession [1]. It can be produced in a variety of ways such as hand-drawn, computer-drawn, model photography, and digitized [2]. The basic principle of animation production is to utilize the characteristics of visual transience of the human eye, by playing a series of still images in rapid succession, causing the audience's visual illusion that the images are moving [3], [4]. In animation production, the application of complex scenes plays an important role in improving the perception, quality, and storytelling of animation works, and with the development of artificial intelligence, it is capable of generating the required content in multi-animation production, including the generation of complex scenes, which not only improves the production efficiency of animation, but also improves its quality and integrity [5]-[8].

The automatic generation of complex scenes has an important role in the whole animation automatic generation system, which will process the scene information obtained after the natural language processing of the story, and carry out the automatic construction of the whole scene according to this information [9]-[11]. When the scene construction is completed, the automatically generated animation can be played in the generated scene [12].

This study proposes an animation complex scene generation method that combines AIGC technology with graph neural network. The study firstly constructs a two-layer graph convolutional network architecture, realizes the hierarchical organization of scene elements through the object layering module, and classifies the objects in the complex scene into primary and secondary objects according to their importance and prominence; secondly, it designs a two-layer feature extraction module, which integrates visual appearance, spatial location and semantic information to form a rich multimodal feature representation; then it establishes a joint training mechanism, which is based on the collaborative learning through the scene graph prediction and region mask optimization to improve the generation quality and generalization ability of the model; finally, a comprehensive evaluation system is

constructed to validate the effectiveness of the method from multiple dimensions such as scene recognition accuracy, generation rate and visual quality.

II. Key technologies

II. A.AIGC technology

AIGC technology [13] is both a class of content produced by AI as a content producer and a way of content production, i.e., a collection of technologies related to the automatic generation of intelligent content. The concept of AIGC possesses a distinction between the narrow sense and the broad sense, with the narrow concept emphasizing more on the content attributes, and the broad concept emphasizing on the technical attributes of AI. From the perspective of content producers, AIGC is roughly understood as a new type of production method that utilizes AI technology to automatically generate content following professional-generated content and user-generated content.

AIGC utilizes AI technologies such as generative adversarial networks and large-scale pre-trained models to generate various forms and styles of content such as text, images, audio, and video by searching for patterns through their own data. From a technical point of view, AIGC is a collection of technologies for automated content generation. According to content categorization, AIGC's technologies can be divided into three main categories: AI-generated natural language technologies, AI-generated visual content technologies, and AI-generated multimodal content technologies. The technology can generate not only texts such as blog posts, program codes, poems and artworks, but also various forms of images and audio-video content.

From the perspective of content producers, AIGC provides them with a new way of generating content, making design more efficient and diverse. From the users' point of view, AIGC will satisfy personalized needs, presenting them with more diverse and accurate information.

II. B.Graph Convolutional Neural Networks

A graph as a data structure is designed to model a set of objects and their interrelationships. Graph Convolutional Neural Network (GNN) [14] is an important application of deep learning in the graph domain, providing a powerful tool for processing and analyzing graph data.

GCN, as an important branch of graph neural network (GNN), is an innovative approach to generalize the concept of convolutional neural network, which is commonly used in image processing, to graph data processing. It has the ability to aggregate the features in the neighborhood of nodes, learn the feature representation of nodes by weighted aggregation, and then perform various prediction tasks.

The working principle of GCN is shown in Fig. 1. Suppose there is a set of graph data containing n nodes, each node has a specific feature vector, the features of all nodes form a $n \times d$-dimensional feature matrix X, and the interrelationships between nodes form a $n \times n$-dimensional adjacency matrix A. The inputs to the GCN model are the feature matrix X and the adjacency matrix A. All nodes are convolved at once to operate as in Eq:

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{1}$$

where $\tilde{A} = A + I$, I is the unit matrix, $\tilde{D}$ is the degree matrix of $\tilde{A}$, computed as $\tilde{D} = \sum \tilde{A}_{ij}$, and H is the matrix of eigenvectors of all the nodes of each layer, and for the input layer, $H^{(0)}$ is equal to X with dimension $n \times d$. The σ is a nonlinear activation function, such as ReLU, and $W^{(l)}$ denotes the trainable parameter matrix for the convolutional transform of the current layer.

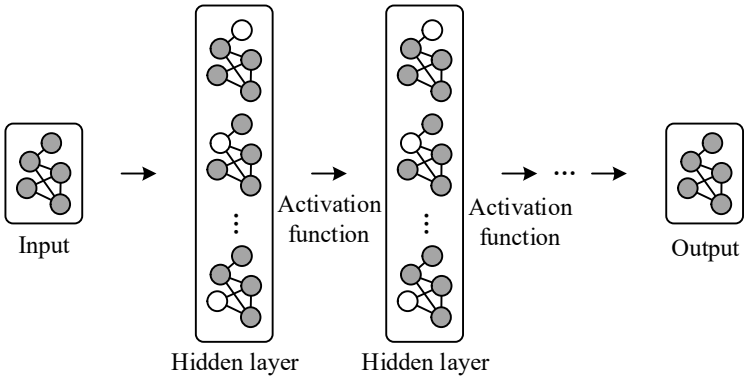

Figure 1: The principle of the GCN

The core function of GCN is to serve as a feature extraction tool that integrates the influence of its neighboring as well as more distant nodes by continuously updating the information of each node in the graph until the state of all nodes reaches stability. In this process, the closer the relationship with a node, the more significant the influence of its neighbors.

III. Complex scene generation model combining AIGC and graph neural networks

III. A. Animation Design Methods Incorporating AIGC Technology

The application of AIGC technology in the generation of complex scenes in animation includes multiple modes and multiple application directions, and is mainly used in the generation of animated scenes. The graph neural network animation method incorporating AIGC technology explored in this study is based on the design method of generating animated scenes with AIGC technology.

In order to investigate the animation design method incorporating AIGC technology, this study uses the content input form (prompt) as the main categorization basis. The animation design methods incorporating AIGC technology are categorized by the type of input prompt, which can be divided into text-based, image-based and video-based animation design methods.

Text-based prompt input generation animation is the simpler design method in the current animation design method integrating AIGC technology, which only requires the designer to input text-based prompt keywords into the AIGC tool to complete the production of images or video animation screen.

III. B. Image scene graph generation method based on two-layer graph convolutional network

III. B. 1) Object Layering Module

For each image I^{SGG} in the animated scene graph dataset D^{SGG}, a series of object bounding boxes $B^{SGG}=\{b_1^{SGG},\ldots,b_{n_b}^{SGG}\}$ are obtained by using the existing target detector. For each object bounding box b_i^{SGG}, the target detector outputs the feature representation v_i of the object and the category label probability distribution o_i.

Meanwhile, the images I^{SGG} from the animated scene graph dataset are input into the salient object detection model pre-trained on D^{SOD} to obtain the segmentation mask $M^{SGG}=\{m_1^{SGG},\ldots,m_{n_m}^{SGG}\}$ for the salient visual information. Where m_i is a separate segmentation mask region. The study uses a VGG-16-based FastenR-CNN as a target detector, and then aligns B^{SGG} with M^{SGG} so as to divide the bounding box in B^{SGG} into the primary object, b_i^{SGG-G}, and the secondary object, b_i^{SGG-S}.

The intersection and merger ratio is a metric used to measure the accuracy of the results of the target detection algorithm, which can measure the degree of fit between the detection results of the detection algorithm and the real labeled bounding box. Referring to the intersection-merger ratio formula, the study defines a pairwise ratio formula:

$$P_{i,j}=\frac{I(b_i^{SGG},m_j^{SGG})}{U(b_i^{SGG},m_j^{SGG})} \tag{2}$$

where $i\in[1,n_b],j\in[1,n_m],I$, I and U denote the intersection operation and the concatenation operation, respectively.

For a segmentation mask region m_j^{SGG}, its corresponding set of object bounding boxes is denoted as S_j^{SGG}. To obtain S_j^{SGG}, the object bounding boxes in B^{SGG} are first sorted from largest to smallest according to the value of pairwise ratios P to obtain the sequence $\{b_1^{SGG},\ldots.b_{n_b}^{SGG}\}$, where for any $1<a<b<n_b$, there is $P_{a,j}<P_{b,j}$. And then, based on this sequence, we get $S_j^{SGG}=\{b_1^{SGG},\ldots.b_t^{SGG}\}$, where $t=\arg\max \mathrm{IoU}(U(b_1^{SGG},..b_t^{SGG}),m_j^{SGG})$. The set of object bounding boxes corresponding to each segmentation mask region collectively constitutes the primary object bounding box set, i.e., $B^{SGG-G}=U(S_1^{SGG},\ldots,S_{n_m}^{SGG})$, and the rest of the object bounding boxes are then categorized as the secondary object bounding box set, i.e., B^{SGG-S}.

III. B. 2) Layered Feature Extraction Module

The inputs to the two-layer feature optimization module are two types of object bounding boxes, namely, the primary object bounding box B^{SGG-G} and the secondary object bounding box B^{SGG-S}. A feature representation of these bounding boxes needs to be generated first. Visual appearance, spatial features and linguistic embedding are considered simultaneously in feature extraction.

(1) Visual features

The feature representation based on segmentation masks is more accurate than the object feature representation obtained based on rectangular bounding boxes. In order to prioritize the main objects, this model uses segmentation masks to generate their visual features. The region mask corresponding to b_i^G is obtained by taking the intersection of the main object bounding box b_i^G with the mask region corresponding to it, i.e., $r_i = b_i \cap m_j (b_i \in S_j)$. Then, the visual features of the main objects are computed as follows:

$$v_i^G = f_{V_1}(r_i) \tag{3}$$

where f_{V_1} is a learning network. As for the secondary object, the object feature representation v_i obtained in the target detection stage is directly used as its visual feature v_i^S.

In addition, the relational features between the primary object b_i^G and the primary object b_j^G are computed as follows:

$$v_{i,j}^{GG} = f_{V_2}((K_i \otimes r_i) \square (K_j \otimes r_j)) \tag{4}$$

where f_{V2} is a learning network, $\otimes$ and $\square$ are convolution and dot product operations, respectively, and K_i and K_j are parameterized by the variances σ_x^2, σ_y^2 as well as the correlations $\rho_{x,y}$ of the Gaussian smooth spatial convolution filter operator. $\sigma_x^2, \sigma_y^2, \rho_{x,y} = f_K(o_i)$, where f_K is a learning network. Similarly, the characterization of the relationship between the primary object b_i^G and the secondary object b_j^S is computed as follows:

$$v_{i,j}^{SG} = f_{V_2}((K_i \otimes r_i) \square v_j^S) \tag{5}$$

(2) Spatial features

For the computation of spatial features, first let $[x_i, y_i, w_i, h_i]$ and $[x_j, y_j, w_j, h_j]$ denote the subject-object bounding box b_i and the object-object bounding box respectively The orientation information of b_j, where (x, y) are the coordinates of the upper-left corner point of the bounding box, and w and h denote the width and height of the bounding box, respectively. The boxdelta is then used to regress b_i to b_j, i.e., $\Delta(b_i, b_j) = \left[\frac{x_i - x_j}{w_j}, \frac{y_i - y_j}{h_j}, \log\frac{w_i}{w_j}, \log\frac{h_i}{h_j}\right]$. Finally, the spatial characterization of the relationship between b_i and b_j is computed as follows:

$$s_{i,j} = f_S([\Delta(b_i, b_j); \Delta(b_i, b_{ij}); \Delta(b_j, b_{ij}); IoU(b_i, b_j); dis(b_i, b_j)]) \tag{6}$$

where b_{ij} is the minimum bounding box containing both b_i and b_j, $dis(b_i, b_j)$ is the normalized distance between b_i and b_j, and f_S is a learning network.

(3) Semantic features

In order to obtain the semantic features of the animated scene graph, a word2vec model is first pre-trained on Wikipedia, and the label o_i of the object b_i is inputted into the pre-trained model to obtain its semantic features l_i. The semantic features of the relationship between b_i and b_j are computed as follows:

$$l_{i,j} = W_l[word2vec(o_i); word2vec(o_j)] \tag{7}$$

where W_l is a projection matrix mapping word vectors into feature space.

When constructing a graph neural network, the features of the nodes are initialized as joint visual and semantic features, i.e., $x_i = [W_v v_i; W_l l_i]$ The features of the edges are initialized as joint visual, spatial and semantic features, i.e., $e_{i,j} = [W_v v_{i,j}; W_s s_{i,j}; W_l l_{i,j}]$. W is the projection matrix used for fusing the features. In addition, a mask head function f_M that maps node features to their corresponding region masks will be learned for optimizing the region masks of the main object.

III. B. 3) Joint training

This method jointly optimizes the predicted scene graphs and the region masks of the main objects. Specifically, the images $I^{SOD} \in D^{SOD}$ from the auxiliary dataset are inputted into the proposed model of the study, and for each main object, its bounding box b_i^{SOD}, initial features x_i^{SOD}, and optimized features $\hat{x}_i^{SOD}$ are obtained. Then x_i^{SOD} is input into the mask function f_M to get the region mask of b_i^{SOD}, i.e. $r_i^{SOD} = f_M(x_i^{SOD})$. The region mask can be optimally updated by the following equation:

$$\hat{r}_i^{SOD} = r_i^{SOD} + f_{ref}(\hat{x}_i^{SOD}) \tag{8}$$

Since the parameters in this formulation exist only in f_{ref}, f_{ref} can be trained with the help of $\hat{r}_i^{SOD}$. Training is performed using a pixel-level cross-entropy loss function, i.e., $L_{seg} = Entropy(\hat{r}_i^{SOD}, y_i^{SOD})$. where y_i^{SOD} is obtained by taking the intersection of b_i^{SOD} and the labeled segmentation mask as the true value at training. The trained mask optimization header function is used to optimize the region mask of the main object with the following optimization formula:

$$\hat{r}_i^{SGG} = r_i^{SGG} + f_{ref}(\hat{x}_i^{SGG}) \tag{9}$$

Then $\hat{r}_i^{SGG}$ is input into Eqs. (3) to (5) to the new feature representation used for subsequent iterative updates. Thus given the datasets D^{SGG} and D^{SOD}, the overall loss function is:

$$L = L_{seg} + L_{opr} + L_{sgg1} + L_{sgg2} \tag{10}$$

III. C. Animated scene image evaluation index

Objective evaluation of images is the most common and widespread evaluation method in the field of computer vision and image processing. The evaluation metrics used in this paper also belong to the full reference category. The four most commonly used full-reference metrics will be introduced next, and PSNR, SSIM and VIF are chosen as the image evaluation metrics in this paper.

(1) Mean Square Error (MSE)

The mean square error belongs to the full reference index based on pixel statistics, which is the sum of the squares of the differences of all the pixels between the original image and the image to be tested, and then find the average value, so as to get the difference of the pixels between the original image and the image to be tested. The specific calculation formula is:

$$MSE = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}[x(i,j) - \tilde{x}(i,j)]^2 \tag{11}$$

where M and N represent the image scale, $x(i,j)$ and $\tilde{x}(i,j)$ denote the pixel values of the original image and the image to be tested at the point (i,j), respectively. the smaller the MSE value indicates the better quality of the image.

(2) Peak Signal-to-Noise Ratio (PSNR)

PSNR is also an evaluation metric based on pixel statistics, which is used to express the ratio between the maximum possible power of a signal and the power of the corrupted noise that affects the fidelity of its representation.PSNR is measured in decibels (dB) in the task of blurred image restoration of a dynamic scene.A larger value of PSNR indicates a better quality of the image.The formula for calculating PSNR is as follows:

$$PSNR = 20\log_{10}(\frac{MAX_I}{\sqrt{MSE}}) \tag{12}$$

where MAX_I is the maximum possible pixel value in the image and MSE is the mean square error.

(3) Structural similarity (SSIM)

Different from MSE and PSNR, SSIM examines the structural differences between two images from three aspects: brightness, contrast, and structure, and belongs to the full-reference evaluation metrics based on structural information.SSIM is based on a sliding window to realize the calculation, similar to the convolution operation. However, the difference is that SSM averages the values of all the windows after traversing the whole image using the window as the final SSIM metric. The calculation formula is shown in Eq:

$$SSIM(x,y)=\frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_2)(\sigma_x^2+\sigma_y^2+c_2)},c_1\sim N_+,c_2\sim N_+ \tag{13}$$

where μ_x and μ_y represent the mean value of the original image x and the image to be tested y, respectively; σ_x^2 and σ_y^2 represent the variances of the original image and the image to be tested, respectively. The larger the value, the better the quality of the image, and if the value is 1, it proves that the evaluated images are all original images.

(4) Visual signal fidelity (VIF)

VIF utilizes the source model, distortion model, and HVS model to obtain feature information for comparison and to obtain the final evaluation results. Among them, the source model is a Gaussian scalar mixture model, which simulates the multi-channel characteristics of the human visual system by statistically modeling the wavelet coefficients of the image pyramid decomposition. The distortion model models the distortion process using a combination of gain and additive noise. The HVS model, on the other hand, represents the human's own distortion channel model. The formulas for each model and the entire VIF metric are shown in Eq:

$$VIF=\frac{\sum_{je.sabbands} I(\bar{C}^{N,j};\bar{F}^{N,j}\mid S^{N,j}=s^{N,j})}{\sum_{je.sabbands} I(\bar{C}^{N,j};\bar{E}^{N,j}\mid S^{N,j}=s^{N,j})} \tag{14}$$

The value of VIF is in the range of [0-1], and a larger value of VIF indicates a better image quality. VIF expands the connection between the image and the human eye in terms of information fidelity, but this type of method is not responsive to the structural information of the image.

IV. Experimental design and analysis

IV. A. Animated Scene Generation Dataset

The study has created its own Animated Scene Generation (ASG) dataset, which is dedicated to the generation of panoramic scene graphs by collecting scene images from different animated movies, where targets are localized into the graphs by means of semantic segmentation annotations. The dataset contains a total of 52k well-labeled animated scene images, 154 kinds of targets (including 93 thing classes and 61 background classes), and 70 relation classes with minimal overlap and sufficient semantic coverage. The dataset is divided into training set, test set and validation set according to the ratio of 7:2:1.

IV. B. Experimental details

In order to have a fair comparison with the state-of-the-art methods, the study trains the proposed network on the training set and validates it on the validation set. The model is trained over 30,000 iterations with the batch size set to 12. The learning rate is initialized to 0.0005 and multiplied by 0.95 every 1000 iterations. Hyperparameters are tuned and optimized on the validation set. All experiments were performed on four NVIDIA GeForce RTX 3090 GPUs deployed in the Pytorch framework.

IV. C. Experimental results

IV. C. 1) Complex Scene Graph Classification and Recognition

The performance of this paper's animated scene generation method combining AIGC and graph neural networks is tested on the training set, test set and validation set in the ASG dataset, and compared with the Cycle Consistent Generative Adversarial Network (Cycle GAN) method and Graph Attention Network (GAT) method. The evaluation metrics used are scene graph recognition accuracy, recall and F1 score, and the comparison results of the animated scene graph recognition performance of different methods on the ASG dataset are obtained as shown in Fig. 2.

By comparing the experimental results of each method, it can be seen that the animated scene graph generation algorithm in this paper, which combines AIGC and graph neural network, achieves good results on the training set, test set and validation set. As can be seen from the figure, the precision, recall, and F1 scores of the three methods

are improved as the amount of data decreases. On the validation set, the precision rate, recall rate, and F1 score of this paper's method are 95.94%, 94.59%, and 95.26%, respectively, which are 7.08-8.20 percentage points higher than that of the Cycle GAN method and 6.15-6.17 percentage points higher than that of the GAT method. It shows that the method of combining AIGC and graph neural network proposed in this paper is effective and can accurately classify complex scenes in animation production.

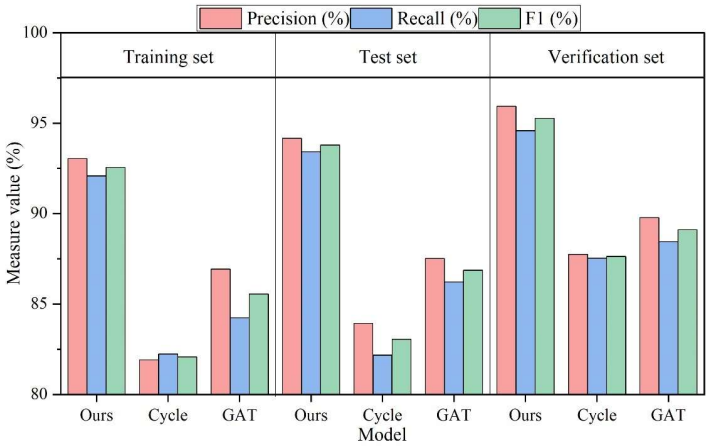

Figure 2: The animation scene diagram identifies performance comparison results

IV. C. 2) Animated scene graph generation rate

In order to further verify the effectiveness of the method in this paper, comparison experiments are carried out, Cycle GAN method and GAT method are used as comparison methods, and the automatic generation rate of animated scene images is used as the evaluation index. Using MAT-LAB to build the simulation test environment, the results of the automatic generation rate of animated scene images of the three methods are shown in Figure 3.

From the table, it can be seen that with the increase of the number of generated images, the generation rate of all methods shows an increasing trend. This may be due to the fact that as the number of images processed increases, the algorithms may have been optimized to improve processing efficiency. The rate of this paper's method when generating 10 images is 4.51 min/each, while the rates of Cycle GAN method and GAT method are 6.44 min/each and 7.21 min/each, respectively. The GAT method has the lowest rate when generating fewer images, and as the number of images increases, the method in this paper always maintains the fastest generation rate. For example, when generating 100 images, the rate of this paper's method is 13.78min/each, while the rates of Cycle GAN method and GAT method are 21.95min/each and 24.35min/each, respectively. Comparing the three methods, it can be found that this paper's method demonstrates a higher efficiency for all tested generation numbers. This result indicates that the method in this paper has a significant advantage in processing efficiency and is more suitable for application scenarios that require rapid generation of a large number of complex animated scene images.

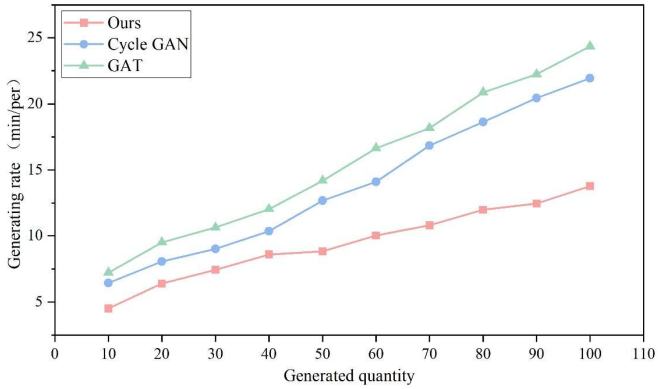

Figure 3: Automatic generation rate of scene image

IV. C. 3) De-blurring effect of processing

In order to validate the effectiveness of the proposed AIGC combined with graph neural network model for deblurring processing, the proposed method is compared with the Cycle GAN method and GAT in the training set, test set,

and validation set of the ASG dataset. The effectiveness of deblurring processing is evaluated using peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and visual signal fidelity (VIF), and Table 1 shows the results of quantitative comparison of scene deblurring processing of different methods.

From the table, it can be seen that the performance of the deblurring processing of this paper's method is better than the other methods, with PSNR, SSIM and VIF of 33.81dB and 0.854, 0.663 on the training set.On the test set, significant results of 34.62dB, 0.871 and 0.692 are obtained. And on the validation set, PSNR, SSIM and VIF are further improved with 35.44, 0.923 and 0.751 respectively.This is mainly due to the fact that: the method in this paper introduces a two-layer feature extraction module, which can better fuse the different scale features of the different phases (i.e., the encoder and the decoder), and hence its performance is better.

Table 1: Different methods scenario to blur the quantitative comparison results

Animation scene	Model	PSNR (dB)	SSIM	VIF
Training set	Ours	33.81	0.854	0.663
	Cycle GAN	31.33	0.766	0.624
	GAT	30.48	0.782	0.626
Test set	Ours	34.62	0.871	0.692
	Cycle GAN	29.73	0.801	0.623
	GAT	30.51	0.828	0.608
Verification set	Ours	35.44	0.923	0.751
	Cycle GAN	31.74	0.831	0.570
	GAT	31.93	0.798	0.619

IV. C. 4) Quality Assessment of Animated Scene Generation

In order to validate the generation quality of complex scene graphs in animation production by fusing AIGC and graph neural network methods, this section uses the validation set in the ASG dataset to conduct a comparison test of the generation quality of complex scenes in animation with the Cycle GAN method and the GAT method. According to the types of complex scenes in the dataset, representative action scenes, urban scenes, natural phenomenon scenes, and sci-fi scenes are selected as research objects. Peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and visual signal fidelity (VIF) were used for the scene graph generation quality. Table 2 shows the evaluation scores of the three methods for generating animated complex scenes.

As can be seen from the evaluation results in the table, the PSNR values of this paper's method for generating action scenes, urban scenes, natural phenomena scenes, and sci-fi scenes are between 31 and 34, which are 1.61-4.45 higher than those of the comparison methods.In the above animated scenes, the SSIM values of this paper's method are 0.859, 0.832, 0.842, and 0.839, respectively, whereas none of the comparison methods' measurements are more In terms of VIF index, this paper's method, compared with the comparison method, improves 6.12%~13.42%. The experiments show that the animated scene generation method combined with AIGC and graph neural network proposed in this paper has superior generation quality, can better retain the picture information, avoid color distortion, and make the style-converted pictures have better visual effects.

Table 2: 3 methods to generate the evaluation of complex scenes of animation

Animation scene	Model	PSNR (dB)	SSIM	VIF
Action scene	Ours	31.94	0.859	0.618
	Cycle GAN	29.56	0.751	0.567
	GAT	29.31	0.774	0.574
Urban scene	Ours	32.68	0.832	0.624
	Cycle GAN	28.59	0.763	0.588
	GAT	29.91	0.774	0.568
Natural phenomenon scene	Ours	31.45	0.842	0.621
	Cycle GAN	29.66	0.762	0.553
	GAT	29.84	0.798	0.584
Science fiction	Ours	33.02	0.839	0.634
	Cycle GAN	28.57	0.774	0.559
	GAT	29.33	0.759	0.589

V. Conclusion

In this paper, by integrating AIGC technology and graph neural network, we have successfully constructed a complex scene generation and dynamic processing framework for animation production. The two-layer graph convolutional network architecture effectively solves the deficiencies of traditional methods in dealing with the relational modeling of complex scenes, and the object layering mechanism and multimodal feature fusion strategy significantly improve the accuracy and consistency of scene generation. Experimental validation shows that the proposed method achieves excellent performance in several key performance indicators. In the scene graph recognition task, the method achieves a recall of 94.59% on the validation set and an F1 score of 95.26%, which is 6.17 and 6.15 percentage points higher than the graph attention network method, respectively, and fully verifies the effectiveness of the two-layer network structure. In terms of generation efficiency, the generation rate is only 4.51min/pc when processing 10 images, which is significantly better than the traditional method, providing efficient technical support for actual animation production. The quality assessment results show that the structural similarity index of the method in the action scene generation task reaches 0.859, and the visual signal fidelity is 0.618, indicating that the generated animation scene has good visual fidelity while maintaining the original image structure information. The technical framework not only provides intelligent tools for animation production, but also provides new technical ideas in the field of computer graphics and artificial intelligence content generation, which has broad application prospects and promotion value.

References

[1] Talabbaev, R. (2024). ADVANCEMENTS IN ANIMATION PRODUCTION TECHNIQUES: A COMPREHENSIVE OVERVIEW. Science and innovation, 3(C5), 24-27.

[2] Park, H. (2021). A study on the improvement of 3D animation production productivity. Journal of Software Assessment and Valuation, 17(2), 101-107.

[3] Saputra, D. I. S., Manongga, D., & Hendry, H. (2021, November). Animation as a creative industry: State of the art. In 2021 IEEE 5th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE) (pp. 6-11). IEEE.

[4] Mou, T. Y. (2015). Creative story design method in animation production pipeline. In DS79: Proceedings of The Third International Conference on Design Creativity, Indian Institute of Science, Bangalore.

[5] Huang, X. Y. (2016, May). Explore scene atmosphere in animation creation. In 2016 International Conference on Applied System Innovation (ICASI) (pp. 1-4). IEEE.

[6] Wang, B., & Wu, B. (2024, July). The Impact of Artificial Intelligence on Chinese Animation. In 5th International Conference on Language, Art and Cultural Exchange (ICLACE 2024) (pp. 153-162). Atlantis Press.

[7] SAAD, R. L. (2024). CHARACTER DESIGN IN AN ANIMATED FILM USING ARTIFICIAL INTELLIGENCE, COMPUTER PROGRAMS OR MANUALLY. Arts and Architecture Journal, 5(2), 98-109.

[8] Stark, L. (2024, June). Animation and Artificial Intelligence. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 1663-1671).

[9] Shen, L., Zhang, Y., Zhang, H., & Wang, Y. (2023). Data player: Automatic generation of data videos with narration-animation interplay. IEEE Transactions on Visualization and Computer Graphics, 30(1), 109-119.

[10] Li, Y. (2021). Film and TV animation production based on artificial intelligence AlphaGd. Mobile Information Systems, 2021(1), 1104248.

[11] Mengya, L., Xiangning, Y., Chi, Z., & Zichu, Y. (2024, December). The Revolution in 3D Animation Film Production Process Under Artificial Intelligence Technology. In 2024 IEEE Smart World Congress (SWC) (pp. 104-109). IEEE.

[12] Zhang, L. (2019). Application research of automatic generation technology for 3D animation based on UE4 engine in marine animation. Journal of Coastal Research, 93(SI), 652-658.

[13] Guoying Chen,Xiaofeng Lan,Kai Liu & Can Cheng. (2025). Research on fusion generation algorithm of visual communication and product design based on AIGC technology. Systems and Soft Computing,7,200237-200237.

[14] Zhengfang He. (2025). Text similarity based on two independent channels: Siamese Convolutional Neural Networks and Siamese Recurrent Neural Networks. Neurocomputing,643,130355-130355.