International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 4617-4627

https://doi.org/10.70517/ijhsa464305

A Study on Scenario Design and Learning Experience Enhancement for Immersive English Speaking Teaching Constructed by Cross-modal Generative Adversarial Networks

Jing Li[1,*]
[1] School of Foreign Languages, Wuhan College of Arts and Science, Wuhan, Hubei, 430345, China
Corresponding authors: (e-mail: 15717159589@163.com).

Abstract The rapid development of artificial intelligence technology has brought new opportunities to the field of education. Aiming at the problems of single scene and poor learning experience in traditional spoken English teaching, this paper proposes an immersive spoken English teaching scene design method based on cross-modal generative adversarial network. By constructing the SPSceneGAN model, the encoder-decoder structure and spectral regularization technique are used to realize the automatic generation of high-quality teaching scenes. The model is trained on the spoken English teaching dataset, which contains 7000 training images and 3000 test images. Experimental results show that the SPSceneGAN model significantly outperforms traditional methods in scene generation quality, with a PSNR value of 38.729dB, an SSIM value of 0.984, and an image processing speed of only 3.921s at a batch size of 500. User testing verifies the effectiveness of the system, with 500 college and university students taking part in a 50-minute comparative experiment, which shows that Students using the immersive teaching scenarios produced significant gains in all three dimensions of prior knowledge level, intrinsic motivation and self-efficacy. The method can effectively enhance students' oral English learning experience and provide a new technological path for intelligent language teaching.

Index Terms Cross-modal generative adversarial networks, immersive teaching scenarios, spoken English teaching, learning experience enhancement, spectral regularization, SPSceneGAN

I. Introduction

The purpose of teaching English as a foreign language is to develop students' oral expression so that they can converse appropriately about general situations in daily life and can carry out conversations or make coherent speeches about materials they have heard or familiar topics [1], [2]. Oral expression must conform to the norms of grammar, phonetics, intonation and the use of appropriate vocabulary [3]. However, most students have the problem of poor listening and speaking skills in English, especially listening, and the important reason is that current teaching cannot provide a practical environment for oral communication and practice [4]-[6]. Therefore, in order to cope with this problem and improve students' English speaking ability, immersive English speaking teaching scenarios constructed by cross-modal generative adversarial networks supported by artificial intelligence are gradually applied [7]-[9].

Immersion teaching method emphasizes that students learn in an authentic and contextualized environment, and through wholehearted participation and experience, students' learning interest and effect are improved [10], [11]. Immersion pedagogy is a student-centered, contextualized, and experiential teaching method [12], [13]. Its core idea is to place students in an authentic and vivid teaching environment, so that students can learn, experience and explore in the context, thus improving students' learning effect [14], [15]. Immersion teaching method is characterized by authenticity, context, experience, and full participation, etc. Teachers can choose a suitable theme and context according to the teaching objectives and students' characteristics, which is of great significance in improving students' spoken English [16]-[18].

This study proposes an innovative approach to construct an immersive English speaking teaching scene using cross-modal generative adversarial network technology. First, the stability and quality of scene generation are improved by improving the traditional GAN architecture, introducing spectral regularization technology, and designing the SPSceneGAN model. Second, a specialized dataset of spoken English teaching scenes is constructed to provide high-quality sample data for model training. Then, the encoder-decoder structure is designed to realize the intelligent conversion from input scenes to target teaching scenes. Finally, the teaching effect of the system is verified through large-scale user testing to assess the effectiveness of the technical solution from multiple dimensions of learning experience, providing theoretical basis and practical guidance for intelligent language teaching.

II. Instructional scenario design based on cross-modal generative adversarial networks

II. A.Generating Adversarial Networks

Generative Adversarial Networks have both a wide range of application scenarios and stability as well as room for enhancement on each side, and are usually composed of two modules: generative models and discriminative models [19], [20]. Inspired by the game principle, the generative model generates samples and the discriminant model evaluates them. The evaluation results of the discriminator are used to adjust the parameters of the generator. Conversely, the samples generated by the generator combined with some real samples are used to tune the discriminator. The goal of our adversarial training is until the discriminator is unable to discriminate the generator's output as true or false, i.e., the generated samples are at the level of being false or true. In the original GAN theory, the generator and discriminator are not required to be neural networks, but only need to be able to fit the corresponding generator and discriminator function. However, in practice, deep neural networks are generally used as generators and discriminators. Because this can solve some of the problems that can not formulate the evaluation function for quantization. Than the immersive English speaking teaching scene design of the good, bad and ugly, is essentially a problem that can not be quantified. Because the model has two deep model components, we need to pay attention to the method of model training, otherwise the output may be unsatisfactory due to various reasons in the training process.

II. A. 1) Discriminators

Most of the discriminators of generative adversarial networks are a type of categorization network and deal with binary classification problems of true and false classification. There are also regressive networks. But the role is the same. That is, the generation of high-dimensional samples mapped to the low-dimensional, the role is equivalent to the generator generated samples with a label, a measure of the generation of samples close to the degree of real samples. For binary discriminators, they are only responsible for answering whether the generated samples are true or false, while the regression problem has a higher dimensionality of the discriminant vectors, which contains more detailed discriminative information and provides a more detailed gradient. For today's large-scale application of GAN in image generation, the discriminator part mostly uses convolutional neural networks. Convolutional neural networks perform convolutional operations on all pixels in an image by defining various types of convolutional kernels. This leads to the extraction of various aspects of features in the image. When we use deep convolutional network as a discriminator for deep adversarial network. A fully connected network is also connected behind it so that the output is a vector available for classification decisions.

II. A. 2) Generator

The generator of a generative adversarial network does not in principle have a fixed structure; it is generally a deep model that maps from low to high dimensions. Its goal is to learn the distribution of real samples from the experience provided by the discriminator, and to generate a sample close to the real distribution from a random Gaussian noise vector. A more representative premise for today's numerous image application scenarios is the use of an inverse convolution structure. Inverse convolution is a mainstream up-sampling process that can scale up small-sized images, mapping a 3*3 feature map to 5*5 by inverse convolution.

II. A. 3) Optimization Functions for GANs and Training Methods

For ease of reading and presentation in the following presentation the discriminator is written as D and the generator is written as G. In the traditional adversarial generative network GAN the optimization function is defined in the form of equation (1). From this formula we can see the strategy of the model. We need to maximize the loss function of D and minimize the loss function of G. The principle of the game is that when my generator is trained to be optimal, the discriminator, no matter how hard it is trained, can no longer discriminate the difference between the generated data and the real sample. For:

$$\min_G \max_D V(G,D) = E_{x\sim p_{data}}[\log D(x)] + E_{x\sim P_g}[\log(1-D(x))] \tag{1}$$

From the above description we can easily see that the training method of GAN is actually a game process of D and G. If G is fixed, $\max V(G,D)$ is the optimization objective, i.e., to train D so that D will have a better discriminative ability. If D^D is fixed, $\min_G \max_D V(G,D)$ is used as the training objective, i.e., train G, so that the samples generated by G are closer to the distribution of the real samples. From this process we can easily analyze. If you want the network to continuously optimize the parameters, you need G, D asynchronous training. That is, train a D to provide discriminative criteria for the training of G. Then fix the G and use the false samples generated by the G to train the D to the next deeper level of discriminative ability. This alternation achieves the common progress of both G and D. The process of alternating training and can be expressed by the following formula.

When fixing G, train D. For:

$$\max E_{x\sim p(r)}\log(D(x))+E_{x\sim p(g)}\log(1-D(x)) \tag{2}$$

transformed into its smallest form:

$$\min-\left[E_{x\sim p(r)}\log(D(x))+E_{x\sim p(g)}\log(1-D(x))\right] \tag{3}$$

When fixing D, train G. The hyperparameter k can be set to indicate that training k times D and then again G. For:

$$\min Loss_G = E_{x\sim P_g}\left[\log(1-D(x))\right] \tag{4}$$

The convergence state of the network can be analyzed as follows. Let the derivative of $Loss\_D$ with respect to $D(x)$ be 0, i.e., the discriminator is optimally trained to the optimal capacity:

$$-\frac{P_r(x)}{D(x)}+\frac{P_g(x)}{1-D(x)}=0 \tag{5}$$

where $p_r(x)$ and $p_g(x)$ denote the distributions of true and false samples. The simplification yields the optimal discriminator as:

$$D^*(x)=\frac{P_r(x)}{P_r(x)+P_g(x)} \tag{6}$$

It is not difficult to find that if the model effect is optimized, the real sample distribution and the sample distribution generated by G are exactly the same. That is, $p_r(x)=p_g(x)$, at this time the probability given by the optimal discriminator is 0.5. The meaning of this 0.5 is that the discriminator can not distinguish the current data from the real one even after further training, and at this time we consider that the ability of G to generate false samples can be recognized. The generated fake samples reach the point where they are as good as the real ones.

II. B.Spectral regularization based generative algorithm SPSceneGAN

Realizing English speaking teaching scene generation is more difficult than the general image generation task. This task firstly requires the completion of feature extraction and reconstruction of the input scene images. Secondly, in the training phase, the model has to learn the rules and meanings embedded in the spoken English teaching scene, e.g., the generation of English teaching equipment should be in the right place, and the generation of student desks should be on the floor, not in the air or on top of other desks. Individual image feature extraction has achieved great success with the development of convolutional neural networks, and subsequent feature reconstruction requires the involvement of generative models, where GANs have made significant progress. In addition to this, in this paper, we will add a spectral regularization constraint model to the model to further improve the quality of generating English speaking teaching scenarios.

II. B. 1) Overview of Spectrum Regularization

For the problems existing in the original GAN, it is either improved from the objective function or from the constraint limitation level. Among them, the spectral normalization method has achieved the current best results since it was proposed, and in this paper, we will introduce the spectral regularization (SN) method in the specific scene generation model to improve the quality of generating English speaking teaching scenes, which in turn will guarantee the students' sense of learning experience.

The role of spectral regularization in the model will be demonstrated theoretically. In the previous chapter, the use of JSdivergence for primitive GANs was analyzed in depth leading to a number of problems such as vanishing gradients, unstable training, and poor diversity. The JSdivergence used by the primitive GAN cannot accurately measure the distance between the real distribution p_r and the generated distribution p_g, because in reality, it is difficult to have an overlap between p_r and p_g, and at this time, the JSdivergence of the distance between p_r

and p_g turns out to be a constant log2, which leads to the instability of gradients, and it is impossible to continue to training the model. In addition, the generator will tend to generate duplicate samples to avoid being penalized, which can cause the model to crash.

To solve the above problems, WGAN proposes to use the Wasserstein distance (EM) instead of the JS scatter measure to measure the distance between real and generated samples. For:

$$W(P_r,P_g)=\inf_{\gamma\sim\Pi(P_r,P_g)} E_{(x,y)\sim\gamma}\left[\|x-y\|\right] \tag{7}$$

where $\Pi(P_r,P_g)$ is the joint distribution of all possibilities consisting of p_r and p_g . the Wasserstein distance ratio JS

That is, $D\in 1-Lipschitz \Leftrightarrow \|\nabla_x D(x)\|\leq 1$ is satisfied for all x, in other words, a differentiable function is a 1-Lipschitz function if and only if, for any x, the modulus of the gradient is less than or equal to 1. WGAN-GP The expression for the discriminant is:

$$\begin{aligned}V(G,D)\approx \max_{D}\Big\{&E_{x\sim P_r}[D(x)]-E_{x\sim P_g}[D(x)]\\ &-\lambda\int_x \max\left(0,\|\nabla_x D(x)\|-1\right)dx\Big\}\end{aligned} \tag{8}$$

This objective function adds a third conditional term, which counts all terms where the mode of the gradient does not satisfy less than or equal to 1. Assigning a penalty parameter λ to these terms, calculating the penalty value, and accumulating all the penalties.

$$V(G,D)=\max_{D\in 1-Lipschitz}\left\{E_{x\sim P_r}[D(x)]-E_{x\sim P_g}[D(x)]\right\} \tag{9}$$

When this cumulative penalty is large enough, it will drag down the value of $\max\{V(D,G)\}$, eventually leading to such that D is no longer the optimal solution. However, this added condition, being valid for all x, makes the penalty very high and introduces a lot of unnecessary computation, so further changes are made to the added condition. In fact, the region that has a substantial impact on the discriminator is the real penalty to be considered. Since the overall purpose of WGAN is to keep the two distributions P_r and P_g close together, the region located between the two distributions must have a substantial impact on the discriminator, so WGAN-GP narrows down the range of x in the penalty term to $P_{penalty}$, and $P_{penalty}$ is the region between p_r and p_g , so the The objective function of the discriminator is transformed into

$$\begin{aligned}V\left(G,D\right)\approx \max_{D}\Big\{&E_{x\sim P_r}\left[D(x)\right]-E_{x\sim P_g}\left[D(x)\right]\\ &-\lambda E_{x\sim P_{penalty}}\left[\max\left(0,\|\nabla_x D(x)\|-1\right)\right]\Big\}\end{aligned} \tag{10}$$

II. B. 2) Data sets and pre-processing

In order to verify that the spectral regularization GAN network can achieve the generation of high-quality spoken English teaching scenes, a scene dataset named Oral English Teaching (OET), containing two types of images, one is the input domain scene image and the other is the target domain scene image, was collected from the web and produced. In order to minimize the interference of the background, the data collected are scene images with clean backgrounds that do not interfere with the scene. The images were cropped to 1024 × 1024 using Photoshop's batch tool, and then they were resized to 512 × 512.In this paper, the latter 3000 images in the input and target domains were selected as the test set, and all the remaining 7000 scene images were used for the training data, keeping the training data roughly 7:3 with the test data.

II. B. 3) SPSceneGAN network structure

The SPSceneGAN model follows the structure of a general GAN, which consists of two parts: a generator and a discriminator. The goal of the generator is to generate specific scenes that are indistinguishable from real spoken English teaching scenes. The generator first learns a mapping between the input English speaking teaching scene and the target image in the conditional domain, and learns the same mapping through other common latent vectors, and then uses the learned mapping to generate the input English speaking teaching scene image into an English speaking teaching scene that is consistent with the target domain. Specifically, the generator inputs the original spoken English teaching scene image, the target domain image is input as a conditional variable which guides the generation of the generator, and the target domain image is copied at input and stitched with the input image to be fed into the generator. The generator reconstructs the new spoken English teaching scene from the input image and then mimics the target scene to generate a fake spoken English teaching scene. The goal of the discriminator is to distinguish the real scene from the generated scene, which is a binary classifier with output samples that are probabilities of the real scene. In this paper, the fake spoken English teaching scene image and the target spoken English teaching scene image generated by the generator are respectively spliced with the original input spoken English teaching scene image and input into the discriminator, using the back-propagation algorithm, which distinguishes between the real and the fake ones by the discriminator and feeds back to the generator, guiding the generator to generate a more realistic image of the spoken English teaching scene, and at the same time, the discriminator's ability to recognize the real English teaching scene is improved, and both of them mutually Against each other, the final English speaking teaching scene generated by the generator is the same as the real English speaking teaching scene, and the discriminator can not be distinguished, to achieve Nash equilibrium.

II. B. 4) Loss function

In addition, this paper uses the SN method to stabilize the model in both the generator and the discriminator, which is a method of constraining the weights, and it can be simply understood as replacing the loss function (11) in the traditional GAN with the form of (12):

$$\min_{\Theta} imize \frac{1}{K}\sum_{i=1}^{K} L(f_{\Theta}(x_i), y_i) + \lambda \sum (w_i)^2 \tag{11}$$

where the canonical terms are replaced with spectral paradigms:

$$\min_{\Theta} imize \frac{1}{K}\sum_{i=1}^{K} L\left(f_{\Theta}(x_i), y_i\right) + \frac{\lambda}{2}\sum_{l=1}^{L} \sigma\left(W^l\right)^2 \tag{12}$$

And the computation of the spectral paradigm is approximated using power iteration. The W matrix is normalized to the equation where the spectral paradigm is constant equal to 1, which in turn makes the parameter gradient of the whole network constant less than or equal to 1, and ultimately achieves the 1-Lipschitz restriction on the generator and discriminator. The additional processing done by the stochastic gradient descent (SGD) algorithm with spectral regularization compared to the traditional SGD is the normalization of the W matrix: $W^l\left(W^l\right) := W^l \big/ \sigma\left(W^l\right)$. ie:

$$\sigma\left(W^l\right) = \tilde{u}_l^T W^l \tilde{v}_l \tag{13}$$

where u and v are the left and right singular vectors of $\sigma\left(W^l\right)$, respectively.

II. B. 5) SPSceneGAN Algorithm Implementation

Both the generator and the discriminator of SPSceneGAN use a convolutional network structure to extract features. Specifically, the generator includes a convolutional structure that performs down-sampling and an inverse convolutional structure that performs up-sampling, and they form an encoder-decoder, and the discriminator includes a convolutional structure that performs down-sampling. The goal of the encoder is to extract the features of the English speaking teaching scene in the input image to form a correspondence with the features of the target English speaking teaching scene image, and the encoder is chosen as a part of the generator not only because of the need to extract the fine features of the input English speaking teaching scene, but also because of the need to generate a new English speaking teaching scene based on the input image, and the encoder is capable of extracting more features than a general convolutional neural network. more features than normal convolutional neural network, it only performs convolutional operation without pooling, and passes more information to the next layer.SPSceneGAN needs to process the input and output which both contain images of the English speaking teaching scene, and the input image is a part of the output image, and it adopts the structure of the encoder and the decoder to firstly extract the features of the input image and the target English speaking teaching scene, and then it maps the input English speaking teaching scene into an English speaking teaching scene similar to the target. Finally, as the number of training times increases, the generated images of the spoken English teaching scene become more and more realistic, and it is more and more difficult for the discriminator to distinguish the truth of the input image, until the discriminator can not distinguish the truth of the input image, and the whole model reaches the Nash equilibrium.

III. Test and Analysis of Scenario Design Scenarios for Teaching English as a Foreign Language

III. A. Algorithm effectiveness test analysis

III. A. 1) Assessment of indicators

As far as the evaluation of the results of the English speaking teaching scene generation work is concerned, the peak signal-to-noise ratio (PSNR) and structural similarity index (PSNR) evaluation indexes are traditionally utilized to calculate the pixel values of the generated English speaking teaching scene images and the real English speaking teaching scene images, so as to measure the relatively objective comparison results. The peak signal-to-noise ratio (PSNR), structural similarity index (PSNR) calculation formula is shown below:

PSNR is a concept in signal processing used to measure the signal-to-noise ratio of a signal. In the context of quality assessment of spoken English teaching scenarios, PSNR represents the ratio between the maximum possible signal value and the error signal power in an image. For images, this is usually the inverse of the mean square error MSE between the original image and the processed image (e.g., compressed or encrypted image.) The MSE is calculated as follows:

$$MSE=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(I_1(i,j)-I_2(i,j)\right)^2 \tag{14}$$

where I_1 and I_2 make the reference image and the image to be compared, respectively, H and W are the height and width of the image, respectively, and the smaller the value of MSE is, the more similar the two images are expressed. The formula for $PSNR$ is as follows:

$$PSNR=10\cdot\log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{15}$$

where MAX_I is the maximum possible pixel value of the image (for an 8-bit image, this is usually 255), and MSE is the mean-square error between the original image and the processed image. $PSNR$ is in decibels (dB), and higher values indicate that the two images are more similar and the generated spoken English teaching scenario is more realistic.

$SSIM$ is a more complex metric that takes into account not only the brightness, contrast, and structural information of the image, but also the visual perception characteristics of the image. The formula for $SSIM$ is as follows:

$$SSIM(I_1,I_2)=\frac{\left(2\mu_{I_1}\mu_{I_2}+b_1\right)\left(2\sigma_{I_1I_2}+b_2\right)}{\left(\mu_{I_1}^2+\mu_{I_2}^2+b_1\right)\left(\sigma_{I_1}^2+\sigma_{I_2}^2+b_2\right)} \tag{16}$$

where μ_{I_1} and μ_{I_2} are the mean of the images, σ_{I_1} and σ_{I_2} are the standard deviations of the images, $\sigma_{I_2I_1}$ are the covariances of the two images, and b_1 and b_2 are small constants added to avoid a zero denominator. The values of SSIM range from 0 to 1, and the closer the value is to 1, the more similar the two images are.

III. A. 2) Loss value analysis

The network model is completed using the Tensorflow 2.0 framework as well as TensorLayer (an auxiliary library for TensorFlow), in addition to the Numpy library, OpenCV library, OS: Ubuntu 18.04 LTS, Graphics card: Nvidia RTX4080Ti, CPU: Intel Core i7- 9700 K. The initialization weights of all convolutional kernels in the generator network and discriminator network are randomly initialized with a normal distribution with mean 0 standard deviation 0.02 $\omega\sim N$ (0.02), and the bias terms are all initialized with 0.0 (b=0). The learning rate is adaptively decreasing according to the number of training rounds, and the initial learning rate is 0.0005. The weight of the original cyclic consistency loss function in the cyclic loss term is set to 10, and the weight of the new covariance matrix loss function λ is set to 1. Fig. 1 demonstrates a comparison of the loss values of the original GAN network with those of the SPSceneGAN generator structure for both the generator and the discriminators. The left and right graphs show the process of real image to generation and the conversion process of generation to real photo, respectively, where the green and red lines represent the generator and discriminator loss value changes of the PSceneGAN generator structure, while the black and blue lines represent the generator and discriminator loss value changes of the original model. The data in the figure show that the improved model can reduce the generator loss and increase the discriminator loss to a certain extent, i.e., the images generated by the improved generator can better “fool” the discriminator, so that the generated immersive English speaking teaching scenes are of higher quality.

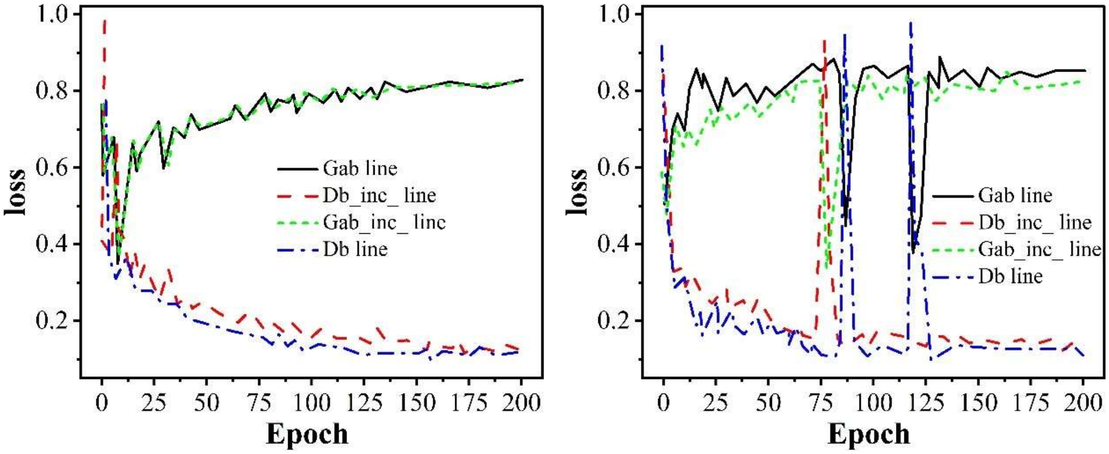
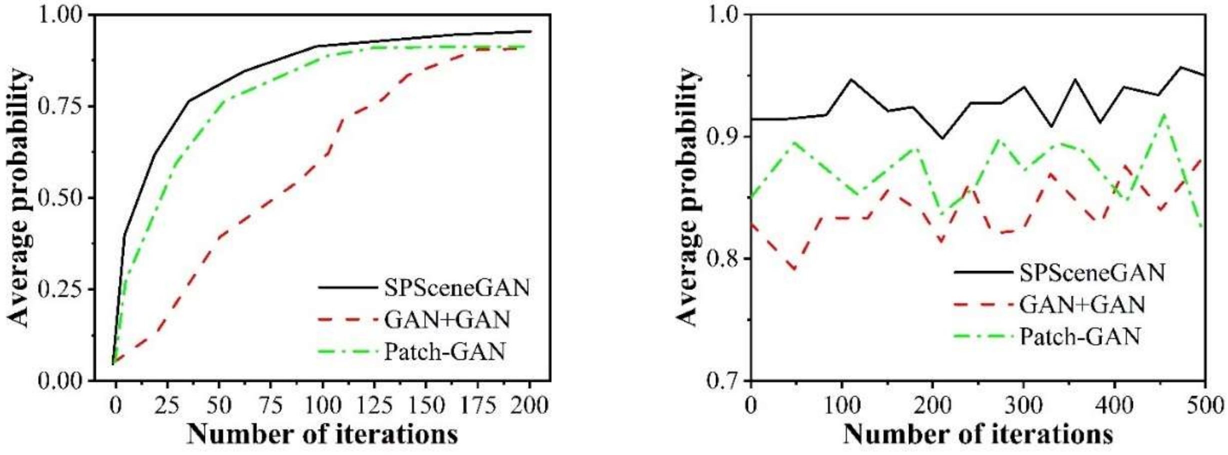


Figure 1: Loss value analysis

III. A. 3) Analysis of experimental results

Two different experiments (Experiment 1 and Experiment 2) were designed to demonstrate the performance of the SPSceneGAN model in English speaking teaching scenarios compared with the current mainstream methods (e.g., multi-GAN overlays). In Experiment 1, the SPSceneGAN model is based on U-Net, Patch-GAN as the basic GAN architecture, and GAN for color conversion of the complete image, and the experimental control group is the unoptimized Patch-GAN model (referred to as the Patch-GAN model) and the general method, the dual-GAN tandem model (referred to as the GAN+GAN model). In Experiment 2, SPSceneGAN is used as the basic architecture of GAN network to complete the immersive English speaking teaching scene generation, and the control group is the RL-GAN model and triple GAN tandem model (referred to as GAN tandem model), and the GAN+GAN model and GAN tandem model are collectively referred to as the GAN overlay model.

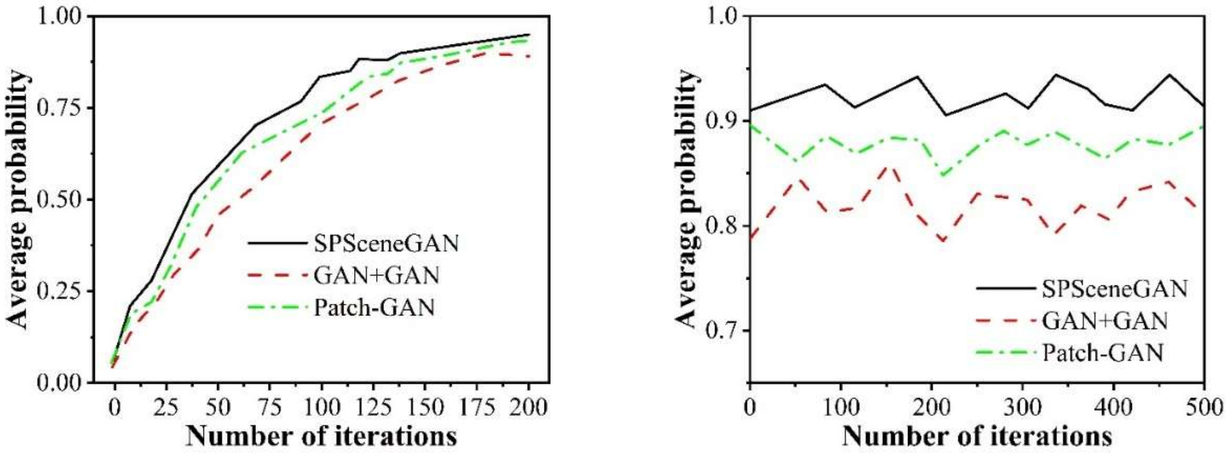
Figures 2(a) and 2(b) show the average probability of the generated images in the pre-trained classifier for different number of iterations of Patch-GAN, SPSceneGAN and GAN+GAN models in Experiment 1, and the classification probability of the output images of the three models in the classifiers for a random 500 inputs, respectively. Similarly, the results of Patch-GAN, SPSceneGAN and GAN tandem model comparison in Experiment 2 are shown in Fig. 3. The results show that as the number of tasks increases, the Patch-GAN model outputs better true probabilities than the GAN stacked model, and the SPSceneGAN model outputs the highest image category truthfulness compared to the Patch-GAN model and the GAN tandem model.



(a)The highest probability

(b)Average probability

Figure 2: Experiment 1 Classification Probability Comparison



(a)The highest probability

(b)Average probability

Figure 3: Experiment 2 Classification Probability Comparison

Table 1 demonstrates the time taken by the above models to process images of different batch sizes. As can be seen from Table 1, the Patch-GAN model image processing speed is better than the GAN overlay model, and the SPSceneGAN model has the fastest processing speed compared to the other two models.

Table 1: Processing time for images of different batch sizes

Model	Batch size				
	100	200	300	400	500
GAN+GAN	0.207s	1.815s	3.106s	4.069s	5.118s
Patch-GAN	0.116s	1.404s	2.229s	3.319s	4.279s
SPSceneGAN	0.089s	1.126s	1.889s	2.978s	3.921s

With the support of the research data, the values of the image evaluation indexes of different models are compared and analyzed, and the comparative analysis of the values of the image evaluation indexes of different models is shown in Table 2. Based on the size of the data in the table, it can be seen that both in Experiment 1 and Experiment 2, compared with the Patch-GAN model and the GAN superposition model, the model in this paper has the highest priority on the immersive spoken English teaching scene, and its corresponding PSNR and SSIM values are 38.729dB (39.11dB) and 0.984 (0.992), respectively, and the PSNR, SSIM The larger the values are, the more the designed immersive spoken English teaching scenarios can meet the students' experiential learning needs.

Table 2: Comparative analysis of Image Evaluation Index values

Model	Experiment 1		Experiment 2	
	PSNR/dB	SSIM	PSNR/dB	SSIM
GAN+GAN	25.71	0.729	24.88	0.753
Patch-GAN	26.78	0.807	25.92	0.824
SPSceneGAN	38.729	0.984	39.11	0.992

III. B. Usability Assessment of English Speaking Teaching Scenarios

In order to test the effectiveness of the system, we implemented a pilot experiment using the traditional way of teaching spoken English as a control group. In this section, it is still describing the user tests and questionnaires and sharing the results of the tests, aiming at English speaking teaching scenarios that can enhance students' learning experience.

III. B. 1) User test participants and objectives

Considering the objectives of the study, students of higher education are in principle optimal as participants. However, the participants should also have the ability to receive questionnaires to give feedback on their psychological changes in the learning process. Considering the age of college students and the feasibility of the research, we chose college students as the subject of the user test. The research in this paper was conducted with voluntary enrollment of students, and the sample consisted of 500 students (250 males and 250 females) who had initially mastered a certain level of English speaking ability. The ages of the children ranged from 18 to 22 years old. Guardian consent was obtained for all data collection and possible presentation. It was also reviewed by the school's privacy and ethics committees. We designed two groups of experiments in which students learned English under different conditions. In Group A, students learned by teaching English in an oral manner. In Group B, students learned through an oral English teaching scenario. Participants were equally divided into two groups based on their pre-tested level of spoken English. Qualitative and quantitative data were collected to analyze the findings. The aim was to investigate whether the participants in the spoken English teaching scenarios had a better learning experience. Both groups were encouraged to perform better in their own research conditions. Invited educational experts were involved throughout the design, implementation and analysis of the results of the user tests.

III. B. 2) User testing process

One week prior to the official start of user testing, students were pretested. The pretest participant questionnaire included measures of a priori knowledge (a priori level of spoken English), intrinsic motivation, and self-efficacy. Participants were divided into two groups based on their level of spoken English as indicated in the pretest to ensure that both groups had approximately the same level of spoken English. The content learned in the user test was identified as something that the participants had not learned. The two groups of participants received 50 minutes of training in different teaching methods. At the end of the learning training, all students received an immediate posttest. We assessed changes in students' learning experiences by calculating pretest and posttest self-reports. Finally, one week after user testing, students were given a delayed retention test to further observe learning effectiveness. A one-week delay in testing was chosen because one week is a typical window for memorization.

III. B. 3) Questionnaire design

The pretest questionnaire contained measures of participants' a priori knowledge level, intrinsic motivation, and self-efficacy. The a priori knowledge measure consisted of 11 questions, the intrinsic motivation measure consisted of 5 questions, and the self-efficacy measure consisted of 5 questions, for a total of 21 questions comprising the questionnaire items. A five-point Richter scale was used, with 1 indicating strong disagreement, 2 indicating disagreement, 3 indicating basic agreement, 4 indicating agreement, and 5 indicating strong agreement. After students' feedback and experts' suggestions, finally the questionnaire has a good reliability performance and can be used for research testing.

III. B. 4) Analysis of experimental results

(1) Intergroup assessment analysis

The assessment results of the learning experience of the two groups of students before the intervention are shown in Table 3, and the assessment results of the learning experience of the two groups of students after the intervention are shown in Table 4. The synthesis of Table 3~4 shows that when there was no intervention, there was no significant difference between the students in Group A and Group B, indicating that the research samples were all in the same, which well ensured the rigor of the research results. After the 50-minute experimental intervention, it was found that the three dimensions of the learning experience of the two groups of students produced significant differences. It can be inferred that the oral English teaching scenario enhances students' learning experience. In order to strengthen the credibility of this result, the within-group assessment analysis will be developed below.

Table 3: Assessment results of learning experience before intervention

Dimension	Group A		Group B		T-Value	P-Value
	Mean	Standard deviation	Mean	Standard deviation		
Prior knowledge level	2.465	0.357	2.481	0.374	1.799	0.112
Intrinsic motivation	2.116	0.296	2.198	0.313	2.187	0.217
Self-efficacy	2.117	0.331	2.343	0.324	2.083	0.296

Table 4: Evaluation results of learning experience after intervention

Dimension	Group A		Group B		T-Value	P-Value
	Mean	Standard deviation	Mean	Standard deviation		
Prior knowledge level	2.852	0.055	3.417	0.449	1.135	0.003
Intrinsic motivation	2.679	0.033	3.528	0.399	1.366	0.004
Self-efficacy	2.754	0.387	3.483	0.378	1.016	0.008

(2) Intra-group assessment

In order to better visualize the differences in the learning experience of students in the two groups, the learning experience of students within the two groups will be assessed and analyzed, and the results of the intra-group assessment are shown in Figure 4, where X1~X3 are the level of a priori knowledge, intrinsic motivation, and self-efficacy, respectively. As can be seen through the data performance in the figure, no significant difference in students' learning experience was produced within Group A, and the significance of the three dimensions of learning experience can be clearly observed within Group B, which is in line with the above findings and further confirms that the English speaking teaching scenarios are able to enhance students' learning experience.

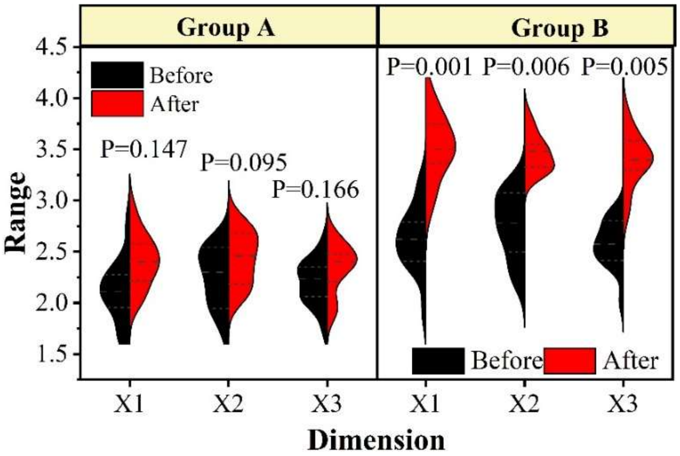


Figure 4: Intra-group assessment results

IV. Conclusion

Spectral Regularization Generative Adversarial Network shows significant advantages in the construction of English speaking teaching scenarios and effectively solves the limitations of traditional teaching modes. The technical validation results show that the SPSceneGAN model obtains a PSNR value of 39.11dB and an SSIM value of 0.992 in Experiment 2, indicating that the generated scene is highly similar to the real scene. Compared with the traditional GAN+GAN model, the processing efficiency is significantly improved, and the processing time is reduced by about 23.4% under the same batch size, which proves the superior performance of the algorithm.

Teaching effectiveness evaluation confirms the positive impact of immersive scenario design on the learning experience. A comparative experiment with 500 college students aged 18-22 found that the experimental group showed significant differences in a priori knowledge level, intrinsic motivation and self-efficacy, and the learning experience was comprehensively improved. The mean value of a priori knowledge water in the experimental group reaches 3.417, which is 19.8% higher than that of 2.852 in the control group, and intrinsic motivation and self-efficacy also reach high levels of 3.528 and 3.483 respectively.

Cross-modal generative adversarial network technology opens up a new development path for English speaking teaching. Through intelligent scene generation, it not only reduces the production cost of teaching resources, but also provides a richer and more diversified learning environment for learners. The technical solution has good scalability and practicality, can adapt to different levels and needs of spoken English teaching scenarios, and provides important technical support and theoretical references for the construction of intelligent language learning system.

Funding

This work was supported by Project of the Hubei Scientific Planning Subject in 2023: "Practical Research on 'AI + VR' Assisting College Oral English Teaching in the 'Intelligence +' Era" (Project Number: 2023GB131).

References

[1] Mahbub, I. S. P., & Hadina, H. (2021). A SYSTEMATIC OVERVIEW OF ISSUES FOR DEVELOPING EFL LEARNERS'ORAL ENGLISH COMMUNICATION SKILLS. Journal of Language and Education, 7(1 (25)), 229-240.

[2] Chen, Z., & Goh, C. (2011). Teaching oral English in higher education: Challenges to EFL teachers. Teaching in higher education, 16(3), 333-345.

[3] Xing, D., & Bolden, B. (2019). Exploring oral English learning motivation in Chinese international students with low oral English proficiency. Journal of International Students, 9(3), 834-855.

[4] Pangket, W. (2019). Oral English proficiency: Factors affecting the learners' development. International Journal of Science and Management Studies, 2(2), 88-98.

[5] Li, H., & Sun, S. (2020). Research on evaluation model of oral English teaching quality based on cloud computing. International Journal of Continuing Engineering Education and Life Long Learning, 30(4), 363-380.

[6] Rahimi, M., & Zhang, L. J. (2015). Exploring non-native English-speaking teachers' cognitions about corrective feedback in teaching English oral communication. System, 55, 111-122.

[7] Li, Y. (2022). Teaching mode of oral English in the age of artificial intelligence. Frontiers in Psychology, 13, 953482.

[8] Wu, R. (2017). A study on SPOC assisted college oral English teaching strategies. Theory and Practice in Language Studies, 7(9), 756.

[9] Wang, J. (2020). Speech recognition of oral English teaching based on deep belief network. International Journal of Emerging Technologies in Learning (Online), 15(10), 100.

[10] Hamilton, D., McKechnie, J., Edgerton, E., & Wilson, C. (2021). Immersive virtual reality as a pedagogical tool in education: a systematic literature review of quantitative learning outcomes and experimental design. Journal of Computers in Education, 8(1), 1-32.

[11] Bhattacharjee, D., Paul, A., Kim, J. H., & Karthigaikumar, P. (2018). An immersive learning model using evolutionary learning. Computers & Electrical Engineering, 65, 236-249.

[12] Wu, C. H., Tang, Y. M., Tsang, Y. P., & Chau, K. Y. (2021). Immersive learning design for technology education: A soft systems methodology. Frontiers in psychology, 12, 745295.

[13] Xie, Y., Liu, Y., Zhang, F., & Zhou, P. (2022). Virtual reality-integrated immersion-based teaching to English language learning outcome. Frontiers in psychology, 12, 767363.

[14] Buragohain, D., Deng, C., Sharma, A., & Chaudhary, S. (2024). The impact of immersive learning on teacher effectiveness: a systematic study. IEEE Access.

[15] Krajčovič, M., Gabajová, G., Matys, M., Furmannová, B., & Dulina, Ľ. (2022). Virtual reality as an immersive teaching aid to enhance the connection between education and practice. Sustainability, 14(15), 9580.

[16] Dengel, A. (2022, May). What is immersive learning?. In 2022 8th international conference of the immersive learning research network (iLRN) (pp. 1-5). IEEE.

[17] Artyukhov, A., Volk, I., Dluhopolskyi, O., Mieszajkina, E., & Myśliwiecka, A. (2023). Immersive university model: a tool to increase higher education competitiveness. Sustainability, 15(10), 7771.

[18] Shi, J., Sitthiworachart, J., & Hong, J. C. (2024). Supporting project-based learning for students' oral English skill and engagement with immersive virtual reality. Education and Information Technologies, 29(11), 14127-14150.

[19] Fei Wu. (2024).English Vocabulary Learning Aid System Using Digital Twin Wasserstein Generative Adversarial Network Optimized With Jelly Fish Optimization Algorithm. Applied Artificial Intelligence,38(1),

[20] Zhao Tong & Song Tianyi. (2022). Establishing a Fusion Model of Attention Mechanism and Generative Adversarial Network to Estimate Students' Attitudes in English Classes. Tehnički vjesnik,29(5),1464-1471.