AniNex – The 6th Workshop on Next Generation Computer Animation Techniques

In conjunction with the 38th International Conference on Computer Animation and Social Agents (CASA 2025)

2–4 June 2025
Strasbourg, France

Accepted Papers

Jian Chang, Bournemouth University, UK
Anil Bas, Bournemouth University, UK
Shihui Guo, Xiamen University, China
Jian Jun Zhang, Bournemouth University, UK

This workshop is partially funded by the EU H2020 Marie Skłodowska-Curie COFUND Scheme (CfACTs 900025) and the Bournemouth University Research Capacity Transformation Fund.

Scene-EEGCNN: Visualization of Zen Meditation Experience Based on EEG-Cultural Heritage Integration

Hui Liang[1] and Longfei Yang[1]

[1] Zhengzhou University of Light Industry, 136 Science Avenue, Zhengzhou 450001, He-nan, China
hliang@zzuli.edu.cn

Abstract. With the deepening of global cultural exchange, Zen culture, as one of the traditional Chinese cultures, has gradually gained admiration from modern society. Zazen, the core practice of Zen, is central to this culture, but the emotional changes of a practitioner during Zazen are difficult to perceive and visualize. By constructing Zen-inspired scenes, the inner world of the practitioner can be depicted. Chinese cultural heritage, including Buddhist sculptures, poetry, and landscape paintings, provides rich materials for presenting these Zen scenes. This paper proposes a method for analyzing emotional changes based on EEG assessment, and maps it to elements of traditional Chinese cultural heritage, using virtual scenes to showcase the emotional fluctuations of the Zazen practitioner. Specifically, this paper introduces the Scene-EEGCNN algorithm, which reads the EEG signals of the practitioner in real-time to assess their emotional state and inner fluctuations. Since the emotional changes of a Zazen practitioner are often difficult for the outside world to detect, this algorithm maps the emotional data to specific elements of Zen culture, constructing a Zen-inspired virtual scene to intuitively represent the practitioner's inner world. With this technology, practitioners can not only gain a deeper understanding of their emotional changes but also share and communicate their Zen meditation experiences with others in a visual way, thus promoting global cultural exchange and understanding.

Keywords: Zen Cultural, Cultural Heritage, EEG, Scene Generation.

1 Introduction

Originating in ancient India, Zen was later transmitted to China, from where it spread to Japan, Korea, and eventually to Western countries such as the United States and various parts of Europe [1]. One of the distinguishing features of Zen Buddhism is its exceptional adaptability; unlike many other religious traditions, it is capable of integrating and absorbing the dominant ideologies of the cultures it encounters. In China, Zen culture assimilated core philosophical elements from Confucianism and Daoism,

evolving into a uniquely Eastern artistic and spiritual tradition. Sitting meditation (Zazen), a central practice in Zen, has gained considerable popularity in modern society as a method of introspection and self-regulation [2]. Contemporary research has demonstrated the positive effects of Zen meditation across a variety of psychological and behavioral domains. For instance, it has been shown to help individuals with addiction or binge-eating disorders regulate cravings[3], improve cognitive and emotional functioning in individuals with attention-deficit/hyperactivity disorder (ADHD) [4], and assist patients with recurrent depressive disorders in managing maladaptive thought patterns and emotional distress [5].

Zen practice, in essence, is a quality of consciousness and is surely a universal and inherent ability of human beings. The visualization of a meditator's inner world has the potential to generate profound impacts across multiple dimensions. On a cultural level, employing EEG to represent the experiential states of Zen practice can transcend linguistic boundaries and foster emotional resonance between civilizations [6]. At the individual level, visualization technologies that track emotional trajectories can offer real-time feedback to optimize meditative practice, while also providing the general public with intuitive access to experiences of inner calm, thereby enhancing psychological resilience [7]. From a societal perspective, such tools could transform approaches to mental health intervention by alleviating collective anxiety and creating spaces for emotional healing through public art initiatives, ultimately strengthening the fabric of social emotional connectedness.

Within Chinese cultural heritage, numerous scenes and artistic forms—refined through nature, humanistic expression, and aesthetic abstraction—profoundly convey the spirit of Zen and reflect the clarity and fluctuations of the meditator's inner state [8]. For example, the winding paths of classical Chinese gardens or the chaotic composition of rocks and trees in traditional landscape paintings may symbolize the mental disarray experienced during meditation. In contrast, the rhythmic motion of a bamboo whisk stirring tea or the repetitive syllables of the Guqin piece Pu'an Incantation can embody a state of focused attention. Meanwhile, the deliberate blank spaces in ink scrolls, the lingering resonance of abruptly ending poetry, or the ephemeral lotus in the palm of a Buddha statue may evoke the meditator's experience of emptiness and inner stillness.

In this study, we adopt a meditative framework and propose the Scene-EEGCNN algorithm to concretely represent emotional changes during Zen meditation. By mapping EEG-based emotional states to culturally grounded Zen-inspired visual scenes, we construct dynamic environments that externalize the inner experiences of meditators. Coupled with an integrated evaluation mechanism, this approach enables the real-time detection and interpretation of emotional shifts. In summary, the contributions of this work are as follows:

- We propose the Scene-EEGCNN algorithm in the form of combining scenes with electroencephalogram (EEG), which makes the inner changes of meditators concrete.
- We constructed a material library of Chinese cultural heritage with Zen connotations, and used EEG to capture the changes in the attention of meditators.

- We classify the emotional changes of the Zazen practitioner captured by EEG using the Scene-EEGCNN algorithm, and map the classification results to a Zen-inspired Chinese cultural heritage database.

2 Related word

While often regarded as a tool for relaxation, Zen meditation is more appropriately understood as a method for cultivating specific emotional states, which can then be used to facilitate self-regulation across multiple functional domains—including physical, emotional, and behavioral processes, as well as interpersonal and intrapersonal relationships. Most therapeutic applications and conceptualizations of Zen meditation emphasize the development of attentional focus and awareness of present-moment experiences, accompanied by a non-judgmental attitude toward these experiences [9]. The primary objective of Zen practice is not merely relaxation, but rather the cultivation of emotional states that foster an accepting mindset. This mindset serves to interrupt habitual patterns of perception and reaction, ultimately promoting clarity of awareness and inner calm.

During the practice of sitting meditation, Zen practitioners typically experience multilayered and dynamic fluctuations in their inner emotional states. Within the rich context of Chinese cultural heritage, numerous scenes and artistic forms—shaped through nature, humanistic philosophy, and aesthetic refinement—effectively convey Zen consciousness and mirror the internal clarity and transformation of the meditator's mind [8].

In the initial stage of meditation, practitioners often struggle with scattered thoughts and emotional instability. This state of mental turbulence can be metaphorically reflected through traditional Chinese garden design techniques such as "winding paths leading to secluded spots" and "borrowed scenery obscured by barriers," which symbolize the confusion and gradual unveiling of the inner self [10]. At this stage, the meditator is akin to a first-time visitor in a garden—unfamiliar with the path, emotionally reactive to external stimuli, and unable to anchor attention.

As the meditation deepens into the second stage, practitioners begin to gain control over their emotional states, with increasing attentional stability and a growing sense of physical and mental harmony. Architectural elements such as pavilions and waterside gazebos within classical gardens are often used as contemplative spaces for self-reflection, representing the psychological transition from external distraction to inward observation [11].

At the most advanced stage, known as the "manifestation of emptiness," the practitioner attains a state of non-attachment and emotional purification, entering a form of awareness that transcends the subject-object dichotomy. Iconic Buddhist statues in Chinese grottoes—such as the meditative Buddha figures in the Mogao Caves of Dunhuang or the serene visage of the Vairocana Buddha at Longmen Grottoes [12]—with their half-closed eyes and tranquil expressions, offer powerful visual metaphors for this ultimate meditative state, representing the unity of consciousness and the transcendence of emotional fluctuation.

At present, it is possible to capture attention changes through EEG signals and classify them, resulting in widely-used datasets, such as the DEAP dataset, SEED dataset, MAHNOB-HCI dataset, Dreamer database, SEED-IV dataset[13], etc. In the SEED dataset, for each emotion, five movie clips with a length of approximately four minutes are selected, which can evoke the desired target emotions. The weight distribution of a trained Deep Belief Network (DBN) is used to select meaningful key channels and frequency bands, and different electrode-set profiles are designed. The experimental results show that the electrode-set pool can achieve relatively stable performance in all experiments across different subjects. Through EEG signals, the user's state is classified into three categories: positive, neutral, and negative. The accuracy of this classification effect is as high as over 80%.

In EEG assessments, the three emotional states—positive, neutral, and negative—correspond to distinct patterns of neural activity: left-right frontal asymmetry, balanced cortical activation, and heightened right-hemispheric cortical activation, respectively [14]. These neural patterns exhibit a profound resonance with the three meditative states observed in contemplative practices: distraction, concentration, and emptiness. Specifically, the distracted state is often associated with elevated beta waves and increased right frontal activity, reflecting anxiety and emotional instability. The concentrated state is typically linked to enhanced frontal midline theta activity, indicating improved emotional regulation and cognitive control [15]. In the state of emptiness, studies have found increased cross-regional synchronization of low-frequency alpha and theta waves, particularly accompanied by decreased activity in the default mode network (DMN), which signifies a diminished sense of self and emotional transcendence [16]. This indicates that the transformation of a Zazen practitioner’s inner state can not only be identified through EEG signals but also aligns closely with the neural mechanisms of emotional regulation, providing a theoretical foundation for mapping subjective Zen meditation experiences to objective physiological indicators.

However, there has not been a method that can visually represent emotional changes in a tangible way, and Zen meditation itself is a form of emotion that is difficult to understand and inherently abstract. This paper proposes the Scene-EEGCNN algorithm, which maps the emotional changes of the Zazen practitioner to scenes from Chinese cultural heritage, enabling the emotional changes to be represented through these scenes. Visualization of emotions can transform them into concrete, perceptible, and visual processes. This approach offers several significant advantages, such as making the emotions easier to understand, facilitating communication, breaking through cognitive boundaries, and assisting in decision-making, among others.

3 Method

3.1 The Mapping of Zen-inspired Scenes to Emotions

In the EEG assessment process, based on the SSED dataset, emotions can be categorized into three types: negative, neutral, and positive, which correspond to three stages of a Zazen practitioner: distraction, concentration, and emptiness. According to these

emotional classifications, we integrate Zen-inspired scenes to represent these different emotional states [17].

As depicted in Figure 1, Scene (a) represents the state of distraction in meditation, which aligns with the negative emotional state. This scene draws on classical design techniques in traditional Chinese gardens, such as "winding paths leading to secluded places" and "obscured and borrowed views" [10], which symbolically mirror the psychological characteristics of this stage. The meandering paths and layered vistas create a spatial sense of ambiguity and disorientation, requiring constant shifts in perspective as one walks—an apt metaphor for the meditator's inner journey through confusion and unrest.

As illustrated in Figure 1, Scene (b) represents the state of concentration in meditation, corresponding to the neutral emotional state. This scene is inspired by elements of traditional Chinese gardens, particularly small pavilions and waterside gazebos, which are often situated at the intersection of movement and stillness, where land meets water. These structures serve to distance the occupant from worldly noise while maintaining a deep connection with the natural flow of the environment [11]. The stillness within the pavilion and the gentle flow of water beyond mirror the meditative process, wherein the practitioner focuses on a single thought and observes the body and mind. This emotional transition—from agitation to calm, from external dependence to internal autonomy—embodies the essence of emotion regulation in the meditative experience [18].

As shown in Figure 1, Scene (c) represents the state of emptiness in meditation, corresponding to the positive emotional state. This scene draws upon the rich visual language of meditative Buddhas in Chinese grotto art [12], such as Amitabha in the Mogao Caves of Dunhuang or the Vairocana Buddha in the Longmen Grottoes—both of which serve as visual metaphors for the manifestation of emptiness. These statues are typically depicted in seated meditation, with half-closed eyes and serene expressions, bodies motionless yet exuding a continuous flow of spiritual energy. They convey a state free of attachment and desire, characterized by selflessness and mental stillness.

Fig. 1. Zen-inspired scenes

3.2 Algorithm Framework

As illustrated in Figure 2. This study proposes the Scene-EEGCNN algorithm, which extracts emotional features of meditators from EEG signals in both time and frequency

domains. These features are then classified using two well-established supervised learning methods: Support Vector Machine (SVM) and Gradient Boosting Decision Tree (GBDT) [19]. SVM is widely utilized in emotion recognition tasks due to its strong generalization ability in high-dimensional feature spaces. GBDT, on the other hand, constructs a powerful ensemble classifier by integrating multiple weak learners (such as decision trees), offering advantages in handling nonlinear features and exhibiting robust performance. These two methods complement each other, enabling effective discriminative analysis of the power spectrum derived from EEG signals.

To enhance the reliability of model evaluation, this study adopts the random subsampling validation strategy. This approach involves repeatedly and randomly partitioning the dataset into training and testing subsets, followed by multiple rounds of model training and evaluation. Such a process effectively mitigates the bias introduced by a single data split, providing a more stable and robust estimate of model performance. Compared to traditional k-fold cross-validation, random sub-sampling offers greater flexibility, particularly in scenarios involving imbalanced data or limited sample sizes—conditions that are especially relevant in this study, which is based on datasets labeled with subjective emotional annotations.

In addition, this study introduces weighted k-Nearest Neighbors (wk-NN) and Logistic Regression as baseline models for horizontal comparison of recognition accuracy, aiming to comprehensively evaluate the effectiveness of the proposed approach. The wk-NN algorithm enhances classification robustness by assigning distance-based weighting factors to neighboring samples, while Logistic Regression is well-suited for both binary and multiclass probabilistic modeling tasks and offers high interpretability. By integrating a Chinese cultural scene library, emotions are mapped to scenes a, b, and c within the database, thereby visualizing the inner world of the Zazen practitioner.

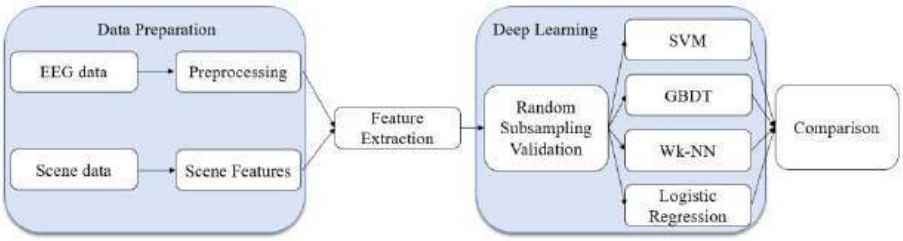


Fig. 2. Algorithm Framework

3.3 Feature Extraction

This work is divided into two parts: the first part focuses on the extraction of emotional features from EEG signals, and the second part involves the mapping of EEG emotional features to scenes.

EEG Feature Extraction. EEG-based emotional feature extraction primarily involves multiple dimensions, including frequency-domain features, time-domain statistical features, and high-dimensional features derived through deep learning techniques. In this study, we calculate features across four frequency bands using a 4-second sliding window with a 2-second overlap. The mean values of the features extracted from these sliding windows are then used as the representative features for each trial. Following this procedure, the total number of features extracted per trial is calculated as (1):

$$(9 \times 32 + 14) \times 4 = 1208 \tag{1}$$

In this experiment, the temporal features extracted from the EEG signals include: peak-to-peak mean, root mean square (RMS) value, and variance. For frequency-domain features, we utilized the Hjorth parameters [20], namely complexity, mobility, and activity, as well as four additional characteristics of the frequency-domain signals, such as maximum power spectral frequency, power spectral density, and power sum. The significance of these features has been clearly analyzed in prior research [21] and was adopted in our experiment. First, the arithmetic mean of the vertical length from the top to the bottom of the time series was calculated, followed by the arithmetic mean of the squared time series, which were determined as the peak-to-peak mean and RMS value, respectively. The next feature is variance, which measures the degree of dispersion in the time series.

The EEG time series are then transformed into the frequency domain using the Fourier Transform, as different frequency bands—such as alpha, beta, theta, and gamma waves—exhibit significant variations across emotional states. In particular, frontal alpha asymmetry (FAA) has been widely recognized as a key physiological marker for distinguishing between positive and negative emotions [14]. Additionally, we compute the total power spectrum and further extract the peak power spectral density along with its corresponding frequency. Three Hjorth parameters, as previously described, are also included in the feature set to enhance the representation of temporal dynamics.

EEG Features and Scene Information Mapping. The garden space constructed with concepts like "curved paths leading to tranquility" and "obscured and borrowed views" possesses greater visual complexity and path uncertainty, which induces exploration pressure and a slight sense of disorientation in individuals. In Scene A, this type of spatial experience is more likely to trigger an increase in δ and θ wave components in EEG, manifesting slight anxiety-related EEG features.[22] From the perspective of emotional modeling, such a scene can be mapped to the negative emotional features extracted from EEG.

A neutral emotional state is typically considered a state of emotional balance. In Scene B, the spatial layout is often open and transparent with soft colors, not directly provoking emotional highs but instead creating a calm and reflective psychological atmosphere.[23] The EEG features in this scene show stable θ waves and moderate α waves, with no significant emotional fluctuation. This neural pattern is highly consistent with a neutral emotional state.

A positive emotional state generally represents an elevation of pleasantness, tranquility, and awareness [24]. In Scene C, the image of the "Zen Buddha" in Chinese grotto art, characterized by a symmetrical, stable seated posture, a gentle and solemn expression, and a serene, peaceful spatial atmosphere, evokes a sense of inner stability and positive emotional response in the observer. The EEG features in this scene show enhanced high β wave activity and increased asymmetry in the prefrontal α waves.

3.4 Scene-EEGCNN

The Scene-EEGCNN algorithm proposed in this paper adopts a model fusion strategy for Support Vector Machine (SVM) and Gradient Boosting Decision Tree (GBDT), combining the prediction results of both GBDT and SVM. Specifically, a GBDT model and an SVM model are trained separately. For new samples, both models make predictions independently, and their results are then integrated using a probability-based weighted averaging method. This combined approach leverages the strengths of both models, thereby enhancing the overall performance and effectiveness of the algorithm.

In the probability-based weighted averaging method suitable for multi-classification, The GBDT model can output the probability of a sample x belonging to each category, denoted as $P_{GBT}(i|x)$, i represents the category; The SVM model can also output the probability of a sample x belonging to each category, denoted as $P_{SVM}(i|x)$, .By assigning weights W_{GBT} and $W_{SVM}(W_{GBT} + W_{SVM} = 1)$ to the two models, the fused probability of the sample x belonging to category m is given by (2). The final predicted category m for the sample x is the one with the highest probability, i.e.: (3), where i iterates over all possible categories.

$$P(m \mid x) = \omega_{GBT} \times P_{GBT}(m \mid x) + P_{\mathrm{SVM}}(m \mid x) \tag{2}$$

$$m = \arg\max(\omega_{GBT} \times P_{GBT}(i \mid x) + \omega_{SVM} \times P_{SVM}(i \mid x)) \tag{3}$$

This paper also employs the weighted K-Nearest Neighbors algorithm (WK-NN) and Logistic Regression for comparison of recognition rates. The two are combined using the Stacking method of model fusion, which can achieve better results in some complex tasks. The Stacking method first divides the dataset into a training set and a test set. The training set is used to train the WK-NN model and the Logistic Regression model respectively.

In WK-NN, the weighted voting formula is used. Suppose there are C categories in total, and the weighted vote number V_j obtained by the j-th category is (4), where ω_i is the weight of the i-th nearest-neighbor sample, and $\mathrm{I}(c_i = \mathrm{j})$ is the indicator function. When the category c_i of the i-th nearest-neighbor sample is equal to j, $\mathrm{I}(c_i = \mathrm{j}) = 1$, otherwise it is 0 . The sample to be classified is assigned to the category with the most weighted votes. In Logistic Regression, a multi-class logistic regression problem is adopted. Suppose there are k categories in total. The true probability $y_k^{(i)}$ that the sample $x^{(i)}$ belongs to the k-th class is 1 only when the sample actually belongs to the k-th class, otherwise it is 0. The probability that the model predicts the sample $x^{(i)}$ belongs to the k-th class is (5). The multi-class cross-entropy loss function is defined as (6).

$$V_j = \sum_{i=1}^{k} \omega_i I(c_i = j) \tag{4}$$

$$\hat{y}_k^{(i)} = \frac{e^{\theta_k^T x^{(i)}}}{\sum_{j=1}^{K} e^{\theta_k^T x^{(i)}}} \tag{5}$$

$$L(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{K} y_k^{(i)} \log(\hat{y}_k^{(i)}) \tag{6}$$

Predictions are made on the test set through these two models to obtain two sets of prediction results. These two sets of prediction results are taken as new features, and then a higher-level model is trained to fuse the prediction results of these two base models. Cross-validation is usually adopted to generate more reliable prediction results as new features, so as to reduce the risk of overfitting and thus improve the accuracy of the comparison.

3.5 Prototype system

Based on the Scene-EEGCNN algorithm, a prototype system was developed, as illustrated in Figure 3. This system utilizes a 32-channel EEG signal acquisition device to collect brain signals, and simultaneously inputs both the EEG data and the cultural scene library into a system powered by the Scene-EEGCNN algorithm.

First, the collected EEG signals undergo data cleansing to eliminate meaningless noise. These cleaned signals are then processed for emotional feature extraction, categorizing the emotional state into three classes: positive, neutral, and negative. This classification allows for the identification of the meditator's current inner state. Using the Scene-EEGCNN algorithm, the system matches the meditator's emotional state with corresponding scenes from the cultural material library, generating a scene that reflects the meditator's internal condition. This visualization enables others to observe changes in the meditator's inner state through the dynamic transformation of the scenes.

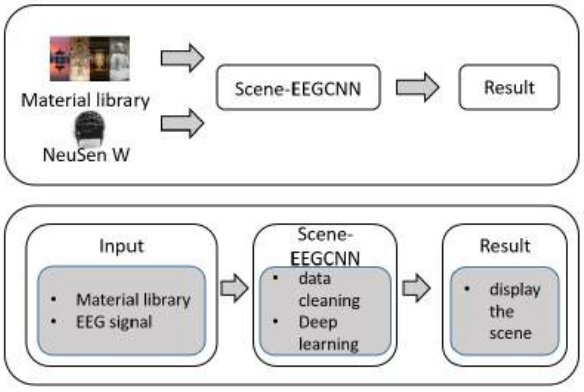


Fig. 3. Prototype system

4 Experiments

4.1 Participants

A total of 50 participants (25 females and 25 males, with an average age of 23 years and 3 months) took part in the study after being fully informed about the nature and purpose of the research and providing their consent. All participants had prior meditation training and were capable of clearly recognizing their emotional changes during the meditation process.

4.2 EEG Signal Acquisition

In this phase, meditation tasks were designed to be both inductive and staged, grounded in the cultural context of Chinese Chan Buddhism. Participants were guided through three sequential psychological states: distraction (restless mind), focus (concentrated breathing meditation), and emptiness (transcendence of emotional fluctuations). Each stage was accompanied by specific audio cues, such as the sound of a wooden fish and verbal instructions, to help participants stably enter the target emotional state. The corresponding time segments were marked and used as reference points for subsequent analysis.

High-precision 32-channel EEG acquisition equipment was used to record participants' brain signals during each meditation stage. The sampling rate was set to 500 Hz, and electrode placement followed the international 10–20 system, with a focus on the frontal, parietal, and central regions. Behavioral labels were recorded simultaneously to facilitate training of the emotion recognition model. For each participant, EEG data were collected with the objective of consistently guiding them into the emptiness state within the meditation process.

4.3 EEG Preprocessing

EEG signals undergo a series of preprocessing steps to remove artifacts. First, a 0.5–45 Hz band-pass filter is applied to eliminate power line interference and DC drift. Then, Independent Component Analysis (ICA) is used to isolate and remove non-neural noise components such as eye movements, electromyographic (EMG), and electrocardiographic (ECG) artifacts. Finally, the cleaned signals are segmented into 4-second sliding windows and labeled according to the corresponding meditation stages, forming a structured, annotated dataset. As shown in Figure 4, EEG signal drift caused by eye movements is a typical example of the artifacts addressed in this process.

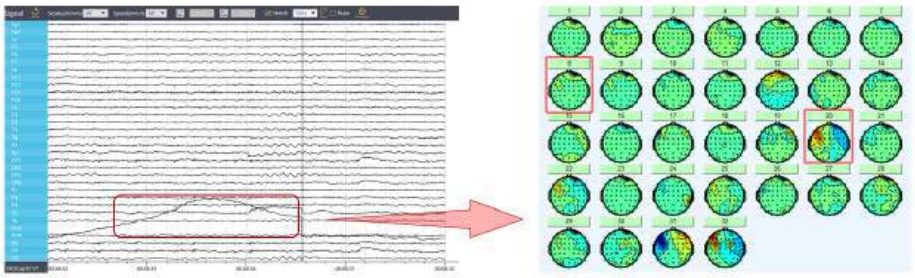


Fig. 4. EEG Preprocessing

4.4 Feature Extraction

Multilevel features are extracted from each EEG signal segment. Time-domain features are derived directly from the raw EEG waveforms and primarily describe signal variations over time. The mean value measures the average potential level within a given time window, reflecting the overall bias of brain activity. Variance indicates the degree of signal fluctuation or activity; higher variance is often associated with heightened arousal states, signifying greater neural instability. Hjorth parameters—Activity, Mobility, and Complexity—further characterize time-domain features by quantifying the signal's power, frequency content, and waveform intricacy, respectively. Frequency-domain features are extracted by applying Fourier Transform and power spectral analysis to the EEG signals, revealing the distribution of energy across different frequency bands. These features reflect the characteristic EEG rhythms associated with various emotional states, providing insights into how neural oscillations differ under positive, neutral, and negative emotional conditions.

4.5 Emotion Classification

In this study, emotions are categorized into three classes: positive, neutral, and negative. Based on the multidimensional features extracted from EEG signals, Support Vector Machine (SVM) and Gradient Boosting Decision Tree (GBDT) are employed for emotion classification. To enhance the stability and reliability of model evaluation, the experiment adopts the Repeated Random Subsampling Validation method. This approach involves repeatedly and randomly splitting the dataset into training and testing sets, with each split undergoing independent training and testing processes. This reduces the risk of incidental errors caused by a single split and provides a more comprehensive assessment of the model's generalization ability across different data subsets. In addition to evaluating the performance of SVM and GBDT, weighted k-Nearest Neighbors (wk-NN) and Logistic Regression are introduced as benchmark models, enabling a comparative analysis of recognition accuracy across different classifiers to validate the effectiveness of the proposed method in emotion classification tasks.

4.6 Experimental Results

To validate the effectiveness of the proposed method in recognizing the three emotional states—positive, neutral, and negative—this experiment performed multidimensional feature extraction on EEG signals collected from participants. Four classification models—SVM, GBDT, wk-NN, and Logistic Regression—were then trained and tested on the extracted features. To minimize the impact of randomness from a single data split, a random sub-sampling validation strategy was adopted: in each iteration, 70% of the data was used for training and 30% for testing, repeated 30 times. The average recognition accuracy across these iterations was used as the final evaluation metric.

As shown in Figure 5, the average recognition accuracies for SVM, GBDT, wk-NN, and Logistic Regression were 85.7%, 84.3%, 78.1%, and 74.5%, respectively, with standard deviations of 2.4%, 2.9%, 3.6%, and 4.2%. Table 1 further illustrates that SVM and GBDT achieved significantly higher classification performance compared to the other models, indicating their superior robustness in handling high-dimensional emotional features. SVM effectively handles nonlinear emotional boundaries through kernel mapping, while GBDT enhances overall performance by integrating multiple weak classifiers, accommodating the nonlinearity and feature interactions inherent in EEG signals.

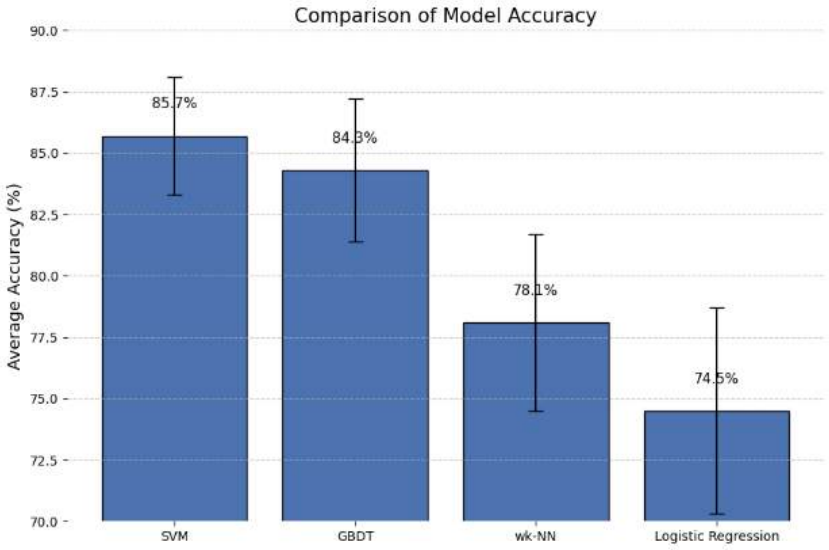


Fig. 5. Comparison of Model Accuracy

A further analysis of recognition accuracy for each of the three emotional categories is presented in Figure 6. The models demonstrated higher classification performance for positive and negative emotions, while the accuracy for neutral emotion was relatively lower. This discrepancy may be attributed to the ambiguous boundaries of the neutral state in subjective experience, which potentially overlaps with both positive and negative emotions, making it more challenging to distinguish accurately.

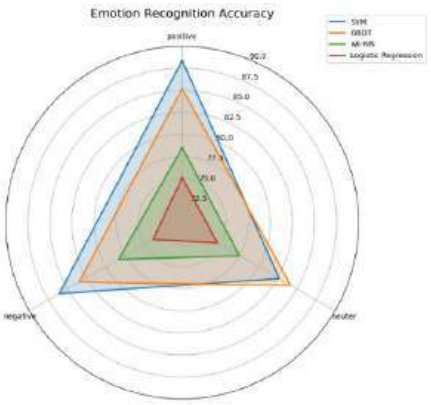


Fig. 6. Emotion Recognition Accuracy

The experimental process is shown in Figure 7. Participants in the experiment experienced a gradual transition from a negative emotional state to a neutral one, ultimately shifting to a positive emotional state. During this process, the emotional fluctuations of participants were monitored in real-time through EEG signals and translated into corresponding changes in virtual scenes. Each image was displayed for 30 seconds to 1 minute, with the system automatically switching scenes based on the emotional changes. Each emotional state corresponded to multiple labels, typically with 5 to 10 images representing negative, neutral, and positive emotions, incorporating elements of Zen culture such as Buddhist sculptures and landscape paintings. The image selection was performed using a predefined algorithm to ensure alignment with the emotional state. Participants could not actively pause or stop the image generation, but could indirectly influence scene transitions by adjusting their emotions, which in turn affected the experimental results.

Fig. 7. Experimental process

5 Discussion

This research is the first to incorporate EEG signals as a key input, integrate them with material library resources, and with the help of the self-developed Scene-EEGCNN algorithm, successfully generate scenes capable of reflecting human inner activities,

opening up a new path for research in this field. Through the generated scenes, we can more intuitively observe and analyze the inner activity patterns mapped by EEG signals, which helps to improve and expand the existing theoretical system regarding the correlation between the brain and psychological activities. In the medical field, this technology is expected to provide new tools for the diagnosis of mental illnesses and the evaluation of treatment effects. In terms of human-computer interaction, it enables systems to understand user needs more accurately and enhance the interaction experience.

Previous research mainly focused on either analyzing the characteristics of EEG signals alone or generating scenes using fixed templates. There was a lack of an effective method to organically combine the two and visualize inner activities.

Although this algorithm can generate scenes with a certain degree of accuracy, when dealing with EEG signals of complex psychological states, there are still cases of scene deviation. Moreover, the computational complexity of the algorithm is relatively high, and its efficiency needs to be improved when processing large-scale data. The collection of EEG signals is affected by factors such as individual differences and the collection environment. The relatively limited sample size may affect the generalization ability of the algorithm.

For future research, it is advisable to consider introducing transfer learning techniques in deep learning to further optimize the Scene-EEGCNN algorithm, thereby enhancing its accuracy and efficiency in complex situations.

6 Conclusions

This research successfully incorporated EEG signals and the material library into the self-developed Scene-EEGCNN algorithm. It innovatively achieved the transformation from EEG signals to scenes in the material library, with the generated scenes capable of representing human inner activities. Through this algorithm, we have built a bridge from the electrophysiological signals of the brain to the visual representation of scenes. This not only provides a new perspective for studying human inner activities but also offers a brand-new approach for technological development in related fields. The experimental results show that Scene-EEGCNN can effectively analyze the psychological features contained in EEG signals and use the material library to generate highly-matching scenes, with an accuracy rate reaching 85.73%.

Acknowledgments. This work is supported in part by the Research Project of Humanities and Social Sciences of the Ministy of Education with grant No. 24YJAZH075, International Cooperation Project of Henan Province with grant No.252102520012，the Research Project of Humanities and Social Sciences of Henan Province with grant No. 2025-ZZJH-370, the Research Project of Intangible Cultural Heritage of Henan Province with grant No. 24HNFY-LX149, the Postgraduate Education Reform and Quality Improvement Project of Henan Province with grant No. YJS2025AL39.

Disclosure of Interests. We declare that we have no fnancial and personal conficts of interests with other people or other organizations that may inappropriately infuence our work. There are

no professional or personal conficts of interests of any nature or any kind in any product, service and/or company that could be construed as infuencing the position presented in, or the review in, the manuscript entitled.

References

1. Chunyuan H.: Research on the Transformation of Confucianism, Buddhism and Taoism Aesthetics in Artistic Practice. Academic Journal of Humanities & Social Sciences 3(1), 81-91, (2020)
2. Shaofeng G, CHENG G, Jing T A O.: Buddhism Initial Dissemination in China: a study of the cross-cultural communication strategies. International Journal of Sino-Western Studies 23, (2022)
3. Kristeller J L, Wolever R Q.: Mindfulness-based eating awareness training for treating binge eating disorder: the conceptual foundation. Eating Disorders and Mindful-ness, 93-105 (2014)
4. Zylowska L, Ackerman D L, Yang M H, et al.: Mindfulness meditation training in adults and adolescents with ADHD: A feasibility study. Journal of attention disor-ders 11(6), 737-746 (2008)
5. Kuyken W, Byford S, Taylor R S, et al.: Mindfulness-based cognitive therapy to pre-vent relapse in recurrent depression. Journal of consulting and clinical psychology 76(6), 966 (2008)
6. Luft C D B, Zioga I, Banissy M J, et al.: Spontaneous visual imagery during medita-tion for creating visual art: an EEG and brain stimulation case study. Frontiers in psychology 10, 210 (2019)
7. Xu Z, Cho Y.: Exploring Artistic Visualization of Physiological Signals for Mindfulness and Relaxation: A Pilot Study. arXiv preprint arXiv 2310, 14343, (2023)
8. PENG Xiuyin, YAO Yi.: The Aesthetic Conception and Ecological Spirit of Chinese Zen Aesthetics. Hundred Schools In Arts 36(3), 17-24 (2020)
9. Kabat-Zinn, Jon.: Mindfulness-based interventions in context: past, present, and future, 144 (2003)
10. Zhang, Zexu.: Winding Paths Leading to Secluded Places: An Environmental Psy-chology Analysis of the Landscaping Art in Classical Chinese Gardens. Modern Horticulture, no. 2, p. 145. doi:10.14051/j.cnki.xdyy (2014)
11. Liu, Xiaoqing.: Application of Zen-Inspired Spaces in the Design of Meditation Archi-tecture in Tourist Attractions. Southwest Jiaotong University, Master's the-sis.Borzab 51-156 (2015)
12. Zhu, Pei.: The Form and Evolution of Buddha Thrones in the Longmen Grottoes. Cultural Relics Appraisal and Appreciation, no. 12, pp. 110–113. doi:10.20005/j.cnki.issn.1674-8697 (2023)
13. Wang X, Ren Y, Luo Z, et al.: Deep learning-based EEG emotion recognition: Current trends and future perspectives. Frontiers in psychology 14, 1126994 (2023)
14. Davidson R J.: What does the prefrontal cortex "do" in affect: perspectives on frontal EEG asymmetry research. Biological psychology 67(1-2), 219-234 (2004)
15. Lomas T, Ivtzan I, Fu C H Y.: A systematic review of the neurophysiology of mind-fulness on EEG oscillations. Neuroscience & Biobehavioral Reviews 57, 401-410 (2015)

16. Josipovic Z.: Neural correlates of nondual awareness in meditation. Annals of the New York Academy of Sciences 1307(1), 9-18 (2014)
17. Morita A, Fukuya I.: Impact of decorative pictures in learning materials: The effect of attention-grabbing features. Applied Cognitive Psychology 37(6), 1352-1365 (2023)
18. Markus H R, Kitayama S.: Culture and the self: Implications for cognition, emotion, and motivation. College student development and academic life. 264-293 (2014)
19. Orima T, Motoyoshi I.: Spatiotemporal cortical dynamics for visual scene processing as revealed by EEG decoding. Frontiers in Neuroscience 17, 1167719 (2023)
20. Hjorth B.: EEG analysis based on time domain properties. Electroencephalography and clinical neurophysiology 29(3), 306-310 (1970)
21. Li X, Song D, Zhang P, et al.: Exploring EEG features in cross-subject emotion recognition. Frontiers in neuroscience 12, 162 (2018)
22. Li R, Ren C, Ge Y, et al.: MTLFuseNet: a novel emotion recognition model based on deep latent feature fusion of EEG signals and multi-task learning. Knowledge-Based Systems 276, 110756 (2023)
23. Imperatori C, Massullo C, De Rossi E, et al.: Exposure to nature is associated with decreased functional connectivity within the distress network: A resting state EEG study. Frontiers in Psychology 14, 1171215 (2023)
24. Zheng W L, Lu B L.: Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Transactions on au-tonomous mental development 7(3), 162-175 (2015)

Exploring the Therapeutic Potential of VR-Based ASMR Animation: A Comparative Study on Relaxation and Sleep Aid

Jiahao Du[1], Lihua You[1][0000−0001−9738−4378], and Jianjun Zhang[1][[0000−0001−9738−4378]

National Centre for Computer AnimationBournemouth University, Fern Barrow, Poole, Dorset, BH12 5BB, United Kingdom
{jdu1, Lyou, jzhang}@bournemouth.ac.uk

Abstract. Although numerous studies have explored relaxation and sleep aid through Autonomous Sensory Meridian Response (ASMR) videos or conventional Virtual Reality (VR) relaxation methods, the integration of VR 3D animation with ASMR and its comparison to traditional VR relaxation methods remains underexplored. To address this gap, this study proposes a VR-based ASMR 3D animation and examines its potential therapeutic benefits in promoting relaxation, aiding sleep, and alleviating stress. First, we investigate a standardized process for creating VR-based ASMR 3D animation games and its impact on triggering the ASMR tingling sensation in VR environments. Then, we develop a VR 3D environment game featuring four different natural environments, along with one ASMR video as a control group. Finally, a comprehensive experiment is conducted to compare the effects of VR-based ASMR 3D animation, conventional VR relaxation, and traditional ASMR videos viewed on a smartphone. Forty seven participants aged 18-35 from Bournemouth University were recruited and divided into three experimental groups. Participants' emotional and physiological responses were monitored using both subjective questionnaires and physiological data collection i. e., heart rate (HR) and electrodermal activity (EDA). Our findings show that VR-based ASMR 3D animation effectively triggers the ASMR tingling experience and offers superior relaxation, sleep assistance, and emotional regulation compared to watching ASMR videos and conventional VR relaxation methods, resulting in a significant reduction in anxiety and stress, as well as increased feelings of calmness and sleepiness. This research highlights the potential of VR-based ASMR 3D animation as a promising tool for relaxation and sleep aid, offering new insights into VRassisted therapeutic interventions.

Keywords: VR-based ASMR · virtual reality · VR relaxation · sleep aid · 3D computer animation · virtual humans · immersive experience.

1 Introduction

Nowadays, anxiety, sleep disorders, and mental health issues have become global challenges, particularly among young people. Insomnia, anxiety, and psycholog-

ical problems are on the rise every year. Non-pharmacological methods such as counsel-ling, yoga practice, travel, and meditation are often used to relieve anxiety and aid sleep in the early stages when people feel stressed and need to relax. In addition to these traditional methods, more and more young people are choosing to watch more convenient and affordable ASMR videos to help with stress relief and sleep.

Previous research has found that the concept of ASMR was first proposed by Jen-nifer Allen in 2010 [1]. It refers to a pleasurable and unique sensation in the cranium, scalp, back, or other body parts triggered by specific visual, auditory, and tactile stimuli such as vision, sound, and touch. These stimuli are known as "triggers" [2]. ASMR has been widely studied not only for its potential to reduce stress and promote sleep [3], but also as a supplementary tool in psychotherapy [4]. ASMR emphasizes immersion and audiovisual stimulation. Traditional ASMR content is mainly viewed through smartphones or computers, which limits the level of immersion it can offer.

With the advancement of VR technology, VR-based ASMR has emerged, offering a more immersive audio-visual experience. It enhances user interactivity and presence, allowing users to quickly enter the virtual environment and more easily triggers the ASMR experience. So far, traditional ASMR research has been extensively explored in fields like psychology and neurology. While the application of ASMR in VR holds great potential, research studies on VR-based ASMR remain limited. In addition, there is a lack of research on how to integrate 3D animation with VR-based ASMR and compare its effectiveness to other VR relaxation methods.

To address these gaps, this paper proposes an approach to VR-based ASMR animated games, as shown in Figure 1. Unlike traditional ASMR videos, which are typically viewed on a 2D screen via a mobile phone or computer, our VR-based approach allows users to interact in a 3D, user-controllable environment using a VR headset. We create a 10-minute character animation in which users

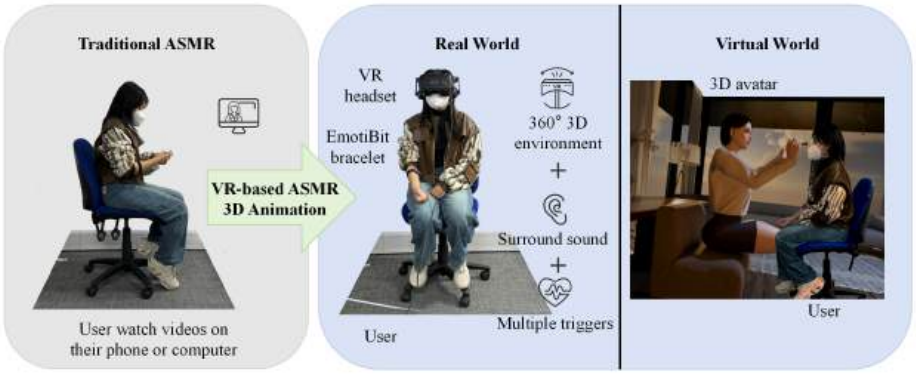


Fig. 1. Comparison of the mode between traditional ASMR and VR-based ASMR 3D animation.

can freely walk around the room to adjust their distance from the character and control the start, pause, and replay of the animation. Additionally, we developed four distinct 3D environments modeled based on conventional VR relaxation mode as a reference group, which allows us to evaluate the effectiveness of VR-based ASMR animation in comparison with conventional VR relaxation environments through comparative experiments.

In our experiments, we used questionnaires similar to the Visual Analog Scale (VAS) [5] and open-ended questions to assess changes in participants' emotional states and their evaluations of different VR environments. We also collected participants' heart rates and electrodermal activity during the experience to assess users' physiological feedback. The contribution of this paper are summarized as follows:

·**A VR-based ASMR interactive animated game.** To achieve this, we integrate (1) motion capture-based animation, (2) UE5 real-time rendering techniques and a 3D VR space, and (3) the construction of blueprints for user controllable animation.

·**A detailed production process of VR-based ASMR 3D animation.** We show the entire production process of VR animation in detail, which provides insights for the subsequent production of a similar type of animation and VR virtual human production.

·**A comprehensive comparative experiment.** Unlike the existing studies, we compared VR-based ASMR animation with conventional VR relaxation environments and non-VR ASMR videos viewed on smartphones, collected questionnaires and physiological data from participants, and analyzed and evaluated the results. The experiment demonstrated that VR-based ASMR animation offers better relaxation and sleep-aiding effects than conventional VR relaxation environments, and is more preferred by users.

2 Related Work

2.1 ASMR

Watching ASMR can alleviate symptoms such as insomnia, anxiety, and clinical depression. Its responses involve various elements such as image guidance, progressive relaxation, meditation, and hypnosis, which can be used to promote relaxation and optimize sleep [6]. Existing research on ASMR mainly focuses on three aspects: (1) exploring the elements (triggers) that induce ASMR sensations, (2) studying the physiological responses of the audience when exposed to ASMR triggers, and (3) using these two aspects as a foundation for research in the disciplines of psychology [7], neurology [8], sociology [9], and digital media [10].

Lochte et al. [11] used nuclear magnetic resonance imaging (fMRI) technology to scan the brains of participants when they were watching ASMR videos, in order to observe which areas of the brain became active. The experiment found that while the viewers were watching the ASMR videos, there was significant

activation of the medial prefrontal cortex (mPFC), which is a brain region associated with social behaviors, self-consciousness, and social cognition. Andersen et al. [12] claim that ASMR represents a form of video-mediated non-standard intimacy. The whispering and touch behaviors that occur in these videos provide the audience with a sense of distant intimacy. They argue that the popularity of ASMR videos can be seen as a reflection of society's need for care, love, and connection. Smith et al. [13] believe that ASMR is not only a sensory response, but also an emotional one, as it utilizes feelings of intimacy and comfort, and embodies social interactions of caring, emotion, and intimacy through the use of visual and auditory elements. The social care and audio-visual performance of these ASMR videos are exactly what is missing in the existing VR relaxation studies.

2.2 VR Relaxation

VR relaxation has been a popular area of VR research and practice in recent years. Several related studies have explored its use as an adjunctive, non-pharmacological treatment in different healthcare programs, with applications in sleep aids [14], anxiety relief [15], chronic pain management [16], and depression treatment [17], offering new approaches to improving mental health and well-being. The most common VR relaxation studies are based on nature scenes, which guide users into relaxation or meditation by recording or designing calm and expansive nature 360° photographs, videos, or VR 3D environments with adapted sound effects. For example, Veling et al. [18] developed VRelax, which includes 360° nature videos and some simple interactive animation elements. Pardini et al. [19] included personalised VR environments in their study, allowing users to select their preferred visual and auditory elements and switch betweenscenes. Cieślik et al. [20] used Japanese garden aesthetics, relaxation techniques, and elements of Eriksonian psychotherapy as the foundation for creating VR 3D environments to alleviate symptoms of depression and anxiety in elderly women. She et al. [21] proposed a VR meditation model based on image transformation and positive feedback. Recently, more innovative VR relaxation experiments have emerged, such as using fractal art images in combination with VR to provide users with a richer and more engaging relaxation experience [22].

These studies and experimental designs are impressive. However, as human beings are social creatures, having a sense of social connection with others can enhance our positive emotions. Compared to conventional natural scenery and meditative environments, few existing VR relaxation studies have incorporated anthropomorphic characters. Even when incorporated, they are typically used only as guides or instructors for users. ASMR videos of role-playing models are known for their immersive blend of visual stimulation and binaural sound, which provides audiences with an immersive viewing experience and virtual social emotions. As with ASMR videos, VR also serves the purpose of providing an immersive experience to users while allowing people to perceive space in a virtual environment, breaking down spatial constraints. Combining VR with

ASMR has the potential to offer a more effective relaxation experience by creating a sense of realism and immersion that is not possible when watching videos on a smartphone or computer.

2.3 VR-based ASMR

Since 2020, Although VR-based ASMR has gained popularity in various life scenarios. VRchat's ASMR rooms have attracted significant user attention, and some of YouTube's 360° or 180° VR-based ASMR videos have more than 10 million views. While ASMR and VR relaxation have garnered significant academic research attention and are widely applied in various therapeutic environments, VR-based ASMR has not yet received enough academic research and experimentation. Existing research focuses more on how VR can be used to better trigger the ASMR experiences. For example, in a study by Aleksandrovich and Gomes [23] on VR multisensory sexual arousal with 140 adult participants, ASMR audio was used for auditory stimulation. Chung et al. [24] invited 53 participants to experience both normal VR stereo audio and ASMR audio in VR to validate that the use of ASMR in VR provides a more immersive experience for participants. Peng et al. [25] proposed a model of asmVR, which uses Unreal Engine to control a 3D character for manipulating the light source to investigate the visual light triggers of VR-based ASMR. Later, they also developed a Unity3D-based multiplayer VR system to enhance users' ASMR experience by integrating wearable devices with vibrotactile feedback [26, 27]. In conclusion, while existing research in this area is emerging, there has been limited exploration into the integration of VR 3D animation with ASMR. Furthermore, current studies lack a detailed demonstration of the production process, an investigation into the system's user experience, monitoring of users' physiological feedback, and comparative experiments on VR-based ASMR.

3 Methodology

Different from previous research and experiments, we provide a detailed overview of the animation game design and experimental processes in Figure 2. Our study outlines the development of a VR-based ASMR 3D animation production process and a comparative experimental framework, resulting in a comprehensive and standardized experimental model of a VR-based ASMR system.

3.1 Outline of VR-based ASMR animation

Existing ASMR videos are generally categorized into two main modes: role-playing and non-role-playing. The non-role-playing mode focuses on the performer's hands, lips, various props, and the sounds they create, such as eating podcasts, chewing, kneading slime clay, and other sounds commonly heard in daily life. In contrast, the role-playing mode of ASMR video involves performers appearing in specific scenes and engaging in role-play interpretation, which

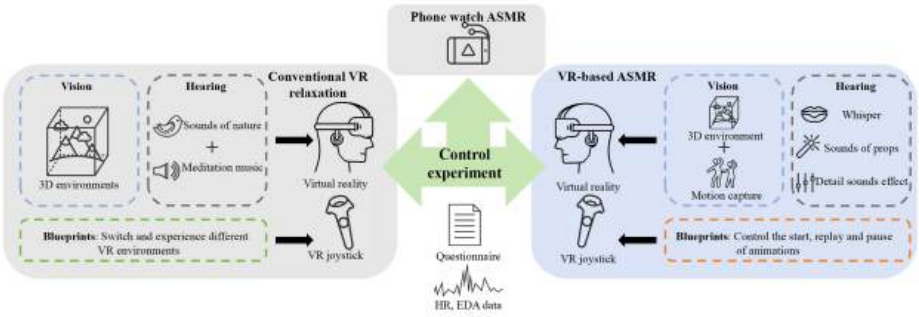


Fig. 2. The detailed process of the control experiment.

constitutes the primary focus of this study. In this type of ASMR video, the performers usually take on roles such as make-up artists, masseuses, medical examiners, or barbers. They simulate physical touch and interactions with the audience in an intimate way, blending visual and auditory elements to provide the audience with a multi-sensory experience.

The VR-based ASMR animation game represents a multimodal synergy, offering an innovative approach to relaxation and sensory engagement. To enhance the sense of immersion for all users, we replace the gender-oriented make-up and haircut performance model with a face massage and light-triggered performance model. This multimodal relaxation model includes the following three primary stimuli:

Visual: the environment used to engage users' visual perspective and the visual frame related gestures and props.

Language: simple and clear command language used to guide users to relax and follow.

Sound: sound resources used to engage users' auditory perspective.

3.2 Design of the Visual Stimuli

Based on our analysis of existing video samples and literature on ASMR, we identified the following performance characteristics of ASMR performers:

Gentle and Non-Aggressive Demeanor. Performers often exhibit gentle and non-aggressive characteristics to promote relaxation and aid sleep. ASMR performers should avoid any aggressive outward appearance and convey a sense of gentleness. Performers often exhibit gentle and non-aggressive characteristics to enhance relaxation and aid sleep. ASMR performers need to lack aggressiveness of the outward appearance with a sense of gentleness.

Gentle body movements. ASMR performances often mimic intimate social interactions, where performers create a closer social distance with the audience. They typically simulate slow stroking and massaging movements, which act as visual 'triggers' to induce a pleasurable tingling or tickling sensation in the audience.

Repetitive Actions. The repetitive nature of the performance, such as simulating repeated stroking of the audience's face, and similar movements can provide a relaxing effect of stabilizing emotions and relieving anxiety. Using these characteristics as a foundation, we developed the necessary steps for our ASMR 3D animation, as outlined in Table 1.

Table 1. The animation process for VR-based ASMR.

Process	Content	Meaning	Results
Introduce (0-1min)	Introduce the content of the game	Users are introduced to VR-based ASMR and introduces role-playing content to give them an idea of the animation next	
Prepare for ASMR (1-2min)	Give the user 45 seconds for the animation to stop	Users are allowed to use the preparation time to observe the 3D VR environment and familiarize themselves with the environment they are in,	
Eye massage (2-3 mins)	The character does a simulated massage of the user's eye area as a trigger	Starting from the basic massage, the user can visually feel the sense of being touched and relax with the character's movements.	
Tuning fork trigger (3-5.5 mins)	The character uses a tuning fork model to close to the user's face and ears	The senses of sight and sound are combined. Tuning fork as it moves closer and further away is synchronized with the movement of the sound for an immersive experience.	
Face massage (5.5-6 mins)	The character does a gentle touch on the user's face to act as a trigger	Repeated massage movements with recorded stroking sound effects further relax the user	
Light pen (6-7 mins)	The character requires the user's vision to follow the movement of light pen	Light trigger is a classic among ASMR's many triggers, when the user focuses on the light pen and moves with it, triggering their sleepiness with hypnotic maneuvers	
Massage ice balls (7-9.5 mins)	The character controls the massage ice balls along the user's cheeks and eyes	The slight sound of water and the movement of the massaging ice balls bobbing close to the face will further calm the user	
End (9.5-10 mins)	The character thanked the user for participating in the experience	Users are encouraged to continue exploring VR-based ASMR with a whisper to end the entire gaming experience	

In addition to animation, the positioning of the player's starting point can effectively convey social distance and perspective. Social distance determines the level of affinity between the animated character and users, just as in the real world, where closer proximity usually means a more intimate relationship. In visual discourse, the choice of close-ups, medium shots, and distant shots suggests relationships ranging from intimate to combative. Perspective, on the other hand, suggests power relationships. When users view the animated character from a

high angle, they are positioned to look down, implying that the character holds less power than the users. Conversely, when users view the animated character from a low angle, they are positioned to look up, suggesting that the character holds more power than the users. When users view the animated character from an eye-level angle, both share equal power. Therefore, the player's birth point should be set at a close-up distance where the user can focus on the character above the waist and at an eye-level angle, whereby the user can maintain an equal relationship with the character, as shown in the screenshot result images of the user's perspective in Table 1.

Visual elements such as specific scene setups and lighting effects can also enhance the ASMR experience. For example, the use of soft lighting and cozy scene setups can create a more relaxing atmosphere. Similar to synesthesia [28], we feel warmth when seeing red, coolness when seeing blue, and a sense of deliciousness when seeing appetizing food. Based on the above discussions, we created a 3D environment of a quiet study with greenery floor-to-ceiling windows and flowers. The natural lighting was adjusted to the dim yellow light of the sunset in UE5 to guide users into relaxation through the visual environment, as shown in Figure 3.

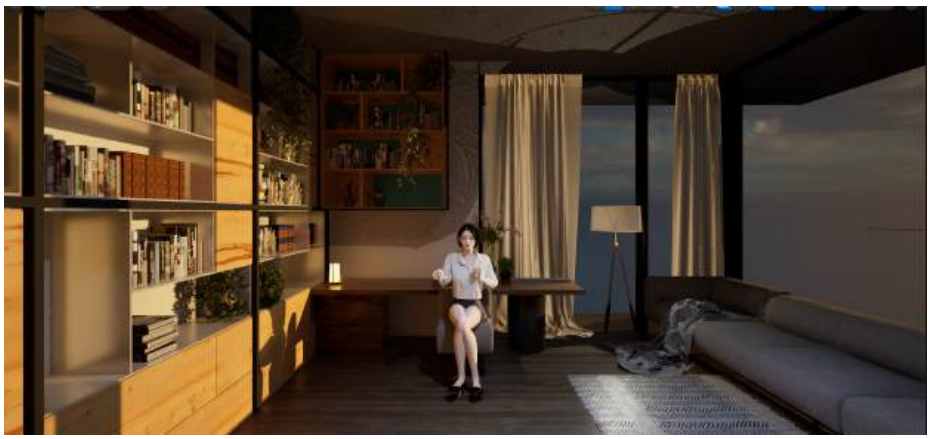


Fig. 3. VR-based ASMR environment.

3.3 Design of the Language and Sounds Stimuli

The main equipment used for sound recording in this study was the TASCAM DR-40X recorder, which captured audio in two categories:

VR-based ASMR sounds. Binaural and 3D surround sound is a major feature of ASMR audio. Performers usually use whisper as one of the "triggers" in ASMR role-playing, by adjusting the distance to the speech microphone and

post-processing sound debugging, they can create an immersive spatial experience through the use of subtle and gentle whispering sounds, usually lip-teeth sounds such as “s”, “sh”, “k”, etc. [29].

Prop Sounds. The sound of props is also an important “trigger”. In this experiment, we recorded the vibration sound of a tuning fork and the water sound produced by the massage puck when massaging the skin. To enhance the richness of the sound effects, we recorded various detailed sounds including the ringing of fingers, the rustling of clothing, and the subtle sound of massaging the skin to recreate the most realistic sound experience. At the end of the recording, we used Adobe Audition (Au) for noise reduction and the Panorama plugin to create a more realistic 3D head surround sound.

3.4 Construction of VR environments and game blueprints

Most VR relaxation studies have used a variety of nature-based virtual environments such as forests, islands, mountains, lakes, waterfalls, and most commonly beaches to promote relaxation.

We used 3D modeling techniques, SpeedTree plant animation, UE5 environment creation, and Quixel Bridge open-source assets to create 4 VR environments, shown in Figure 4. 4 of them are normal VR relaxation environments, including snowy mountains, rivers, lakes, caves, meditation rooms. Based on previous research, we recorded binaural nature sounds such as rivers, birds, water droplets, wind blowing through leaves accompanied by meditation-guided music to assist users to achieve relaxation.

The game blueprints are primarily used to link users with the VR environments. A simple control system is preferred, as it allows users to focus more on

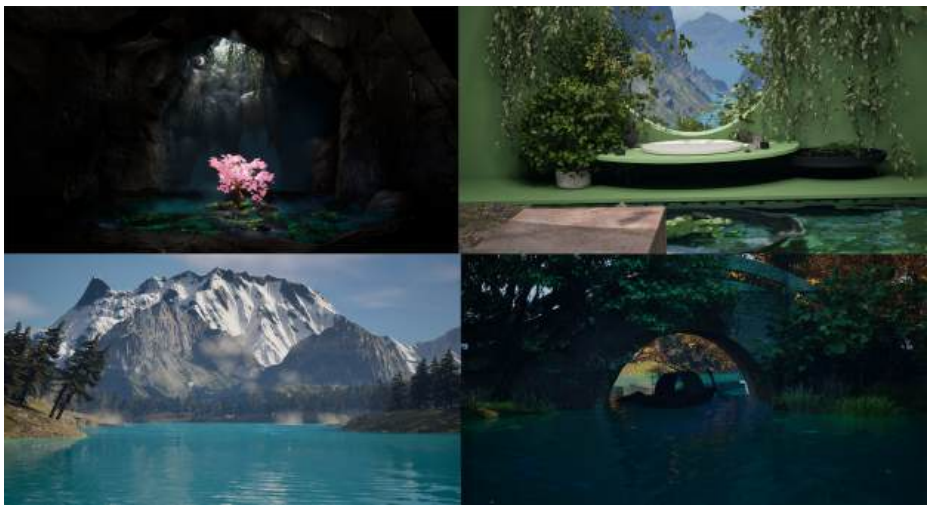


Fig. 4. Conventional VR relaxation environments.

relaxation compared to a complex control system. In our design, users can use VR joystick buttons or computer-specific keys to switch and experience different VR environments, as well as control the start, replay and pause of animations.

3.5 Watch ASMR Videos on Smartphone

An ASMR video was selected and taken from YouTube by the author. The video length is the same as the VR-based ASMR animation game length of 10 minutes, including (1) face massage, (2) light triggers, (3) tuning forks, and (4) whispers and other triggers that are the same as the game.

3.6 Questionnaire Survey

We designed a set of participant questionnaires for this experiment comprising the following sections:

Sample characteristics: The basic information about participants was collected through recording participants' age and gender, their experience and frequency of using VR, watching ASMR videos, and using VR-based ASMR.

Emotion state change measurements: A visual analogue scale (VAS) similar to the one used by Navarro-Haro et al. [5] was used to monitor participants' emotional and cognitive changes before and after the experiments.

Experiment evaluation form: This section contains a list of "trigger" elements similar to those used in ASMR research by Barratt and Davis [30].

Open-ended questions: These questions allow participants to describe their feelings and provide suggestions about the experiment.

3.7 Participant Physiological Data Monitoring

In this study, we used the Emotibit bracelet [31] to monitor participants' heart rate (HR) changes and electrodermal activity (EDA) during a VR-based ASMR 3D anima-tion experiment, a conventional VR relaxation experiment, and a smartphone-based ASMR viewing experiment. EDA is an indicator of physiological arousal during emo-tional, cognitive, and physical behaviours and decreases with physiological relaxation such as sleep or rest [32]. Similarly, changes in HR also represent the level of calm-ness of the participant. Measures of participant HR and EDA changes have been similarly used in several studies of physiological changes in ASMR audiences. For example, Engelbregt et al. [33] examined the effects of ASMR videos on mood, at-tention, HR, EDA, EEG (electroencephalogram), and their interactions with personali-ty factors in 38 young adults. Their study found that in all participants, regardless of whether they felt tingling or not, HR decreased while watching the ASMR videos, suggesting that ASMR is associated with relaxation. Additionally, participants who experienced ASMR-triggered sensations had higher EDA after watching the ASMR video. This increase in EDA levels contradicts the relationship between ASMR and relaxation. However, Poerio et al. [7] also found that ASMR elicits arousal

in individ-uals who experience it, suggesting that ASMR may be associated not only with relax-ation but also with increased arousal. These results suggest that the tingling sensation triggered by ASMR has a physiological basis and is central to the experience itself. Therefore, we used HR data to assess whether participants were relaxed and EDA data to determine whether participants were ASMR-triggered.

4 Experiment

To assess the effectiveness of VR-based ASMR 3D animation compared to conventional VR relaxation in promoting relaxation, aiding sleep, and relieving stress, we conducted a controlled trial by recruiting 47 participants and engaging them in two different VR games and ASMR videos viewed on a smartphone. The experiments were developed in Unreal Engine 5.3.2 as a game program for the Windows platform, and the VR device we chose to use was the HTC Vive Pro 2.

4.1 Participant

A total of 47 Bournemouth University students participated in this experiment, aged between 18 and 35 years. The group consisted of 23 females and 22 males including 23 undergraduates, 11 postgraduates, and 11 PhD students. The participants had varying levels of exposure to VR systems and knowledge of ASMR. Among them, 33 had experience with VR, 10 had used VR for relaxation, 25 had watched ASMR videos, and 3 had experience with VR-based ASMR. All participants provided in-formed consent to participate in the 3 experiments. To ensure the accuracy of the experimental data, the 47 participants completed all three experiments in a random order, with a one-week interval between each successive experiment. Each experi-ment consisted of a five-minute pre-experience questionnaire, a ten-minute VR or mobile phone experience, and a five-minute post-experience questionnaire.

4.2 Procedure

Before the experiment began, the participants completed a form to provide their basic information and an emotional state scale, which was used to quantify the par-ticipants' emotional state. Following this, they were seated in a chair, equipped with an emotibit physiological data measurement bracelet and a VR device. Then, they adjusted the VR device to their preferred viewing angle to ensure comfort and an optimal experience, as shown in Figure 5.

In the control trial of conventional VR relaxation, during the preparatory testing phase of the experiment, we found that when participants were asked to experience four VR relaxation environments within ten minutes, they usually switched quickly between different scenes to experience the novelty of VR, which did not help relaxa-tion. Additionally, some participants indicated that

Fig. 5. A participant experiencing (a) Conventional VR relaxation. (b) VR-based ASMR. (c) ASMR video on smartphone.

experiencing only one environ-ment for ten minutes would be boring. Therefore, the participants were given the op-tion to select their preferred two environments from the four different 3D environ-ments we created. This flexibility allowed participants to tailor the experience to their preferences, ensuring a more personalized and engaging relaxation session. Once everything was set up, the participants could switch between the VR environments and animations by using the VR joystick or specific computer buttons. After each experiment, the participants were asked to complete the Emotional State Scale again to assess their emotional changes before and after the experiment. They also filled out the Experiment Evaluation Form to record their ratings of the experiment. Finally, the participants answered open-ended questions to share their feelings during the experience.

5 Results and Discussion

5.1 Results of the Participant Questionnaire

The Participants were asked to indicate whether they experienced any triggered sen-sations such as pleasurable "tingling" or "numbing" sensations while experiencing the VR-based ASMR 3D animation and ASMR videos by answering with 'yes' or 'no'. The final result showed that 43 out of 47 participants experienced ASMR-triggered sensa-tions in the VR-based ASMR animated game experiment, while 36 out of 47 partici-pants experienced ASMR-triggered sensations in the experiment using a smartphone to watch ASMR videos. The participants who

Table 2. Different triggers and their number of selected times, (a) VR-based ASMR, (b) ASMR video on smartphone. captions should be placed above the tables.

Trigger	Number	Trigger	Number
Light pen	14	Slow movement	10
Whisper	16	Tuning fork (visual & hearing)	30
Personal attention	21	Repeat action	12
Touched feeling	24	Ice ball (visual & hearing)	26

(a)

Trigger	Number	Trigger	Number
Light pen	8	Slow movement	8
Whisper	10	Tuning fork (visual & hearing)	20
Personal attention	16	Repeat action	7
Touched feeling	15		

(b)

answered "yes" were asked an addi-tional question using the asked to select the elements that triggered their sensations, as shown in Table 2. The table shows VR-based ASMR animations have a signifi-cantly higher number of departures than ASMR viewed on smartphones with the same type of "triggers". This means the VR-based ASMR animation game is more effective in triggering the ASMR experience in users compared to watching the ASMR videos on a smartphone.

The section of Emotion State Change Measurements section was used to monitor participants' emotional and cognitive changes before and after the experiment. The "Sleepy" option was added to assess the sleep aid effect of the research program. The participants were asked to complete the same form before and after the experi-ment by rating the following eight emotions on a scale of 1 to 7 (1 = "not feeling at all", 7 = "feeling very much"): sleepy, calm/relaxed, happy, anxious, sad, angry, surprised, and energized/energetic. The data obtained from the three experiments were divided intothree groups, with each group including 47 participants' pre-experience state (Time1) and post-experience state (Time2) from one experiment. Based on the algorithm in [22], the mean value (M) and standard deviation (SD) of the difference between Time2 and Time1 were calculated to obtain the participants' standardized change values across the three experiments, as shown in Figure 6. The "Happiness", "Sleepy", "Calm/relaxed",

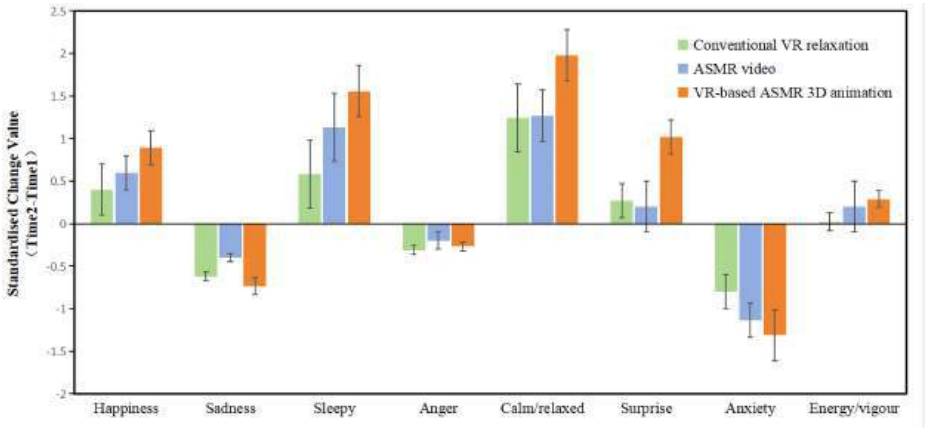


Fig. 6. Standardised change values (T2-T1) for all outcomes as a function of condition.

"Surprise", and "Energy/vigour" scores increased after the experience. The mean value of the "Calm/relaxed" state of the VR-based ASMR animation M(T2-T1) = 1.98 which is 1.59 times higher than conventional VR relaxation and 1.56 times higher than the ASMR video. The mean value of the "Sleepy" state M(T2-T1) = 1.56, which is 2.69 times more than conven-tional VR and 1,37 times than ASMR videos. And the "Sadness", "Anger", and "Anxiety" scores decreased after the experience. The mean value of the "Anxiety" state of the VR-based ASMR animation M(T2- T1) = -1.31, which is 1.64 times lower than conventional VR

relaxation and 1.17 times lower than the ASMR video. This indicates that both the experimental and control groups were effective in relieving anxiety and helping users relax. The results showed that the VR-based ASMR 3D animation led to greater changes in multiple emotional state scores than the conven-tional VR relaxation mode and ASMR videos and has a stronger ability to regulate emotions and promote sleep and relaxation.

After completing the three experiments, participants were asked to choose their favorite of the three experiences. Of the 47 participants, 32 selected the VR-based ASMR animation game, 9 chose the ASMR video on a smartphone, and 6 preferred the conventional VR relaxation. The VR-based ASMR animation game was favored by 68.11

At the end of the experiment, the participants were asked to answer a series of open-ended questions in writing.

Open Question 1 (OQ1): *"Please summarize your overall response to the experi-ence as detailed as possible".*

OQ2: *"Will you be happy to try this type of experience in the future?"* (Answer "yes", "no" or "uncertain" and explain why). This question explored the motivational aspects of the intervention (uptake likelihood)

OQ3: *"If you have done any guided relaxation exercises such as ASMR video meditation or guided breathing outside of VR or with VR in the past, how does this compare to your previous experience?"*

A lot of positive feedback was gathered from the participants in this session. For example, one participant said, *"I have previously watched some ASMR videos online. The VR-based ASMR animation is better than the videos. And this is the first time I've felt relaxed and a bit sleepy through ASMR."* (Participant 6) Another participant said, *"I have social barriers, so avatars that aren't real people make me more re-laxed."*(Participant 13) and *"I'd like to experience more different ASMR games for VR, preferably with the option to personalize the character to my liking."* (Partici-pant 41).

The majority of participants in the VR-based ASMR animated game reported their feelings *"relaxed"*, *"sleepy"*, *"immersive"* and found the experience *"more interesting"* at the end of the experiment.

There were also a small number of participants who expressed different views: *"I prefer natural scenery, and watching VR of natural scenery makes me feel more re-laxed"*, and *"It is more convenient to watch ASMR videos on phones, although VR provides a better sense of immersion, the VR equipment is too heavy!"*

5.2 Results of the Participant Physiological Data

Figure 7 shows the real-time physiological changes recorded by the Emotibit over a 10- second period while participants experienced different VR games and watched a video on a smartphone during the experiment. All experiments resulted in a decrease in HR, indicating a gradual calming of the participant's emotions. Compared to (a) and (c), the HR drop in (b) is more pronounced, which can be interpreted as a better relaxation effect provided by VR-based ASMR during this 10- second period. Paradoxically, however, the EDA data,

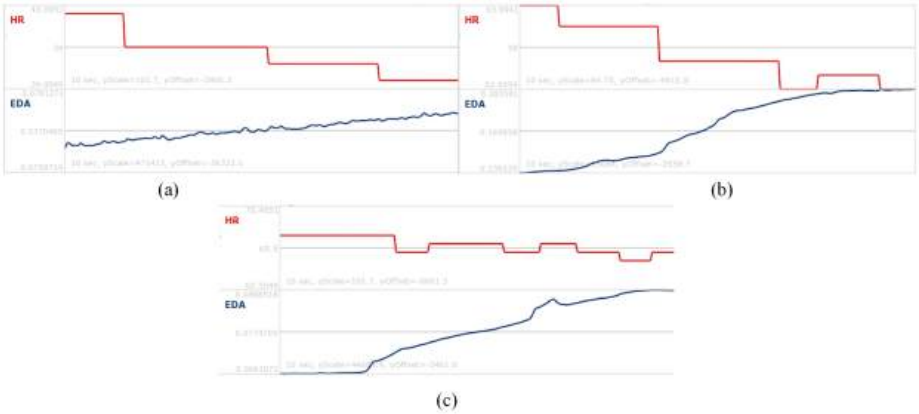


Fig. 7. The participants' HR and EDA data displayed in real-time by the EmotiBit device for (a) conventional VR relaxation, (b) VR-based ASMR, (c) ASMR video on smartphone change values (T2-T1) for all outcomes as a function of condition.

which reflects emotional arousal, showed signifi-cant changes in both the VR-based ASMR experiment and the ASMR video. And compared to ASMR videos, the EDA curve for emotional arousal in VR-based ASMR animations produces changes in a shorter time and is flatter.

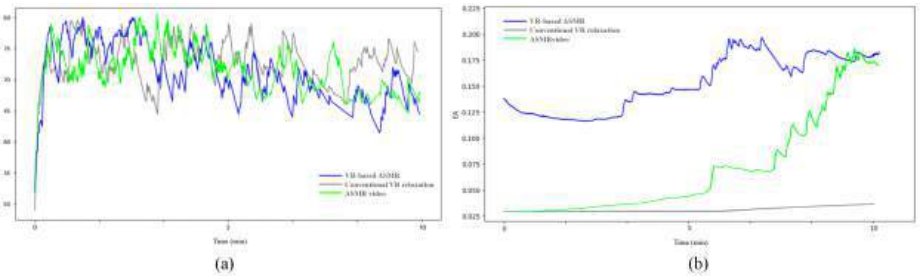


Fig. 8. (a) HR comparison, (b) EDA variability comparison.

Figure 8 (a) shows the one participants' HR changes throughout the experience of the different experiments. Compared with conventional VR relaxation and ASMR video, VR-based ASMR resulted in a more significant decrease in the participants' HR. Together with the results from the previous emotional state change questionnaire, it can be concluded that VR-based ASMR has a better relaxation effect and can help users reach a calm state in a short period of time.

Figure 8 (b) shows the change in EDA for a participant who experiences three different experiments, and it can be seen that when the participant experiences conventional VR relaxation, almost no change in EDA is produced. Combined

with the previous data analysis, it can be concluded that conventional VR relaxation can provide users with a sense of relaxation and immersion but has little effect on their emotional arousal. Compared to ASMR videos, VR-based ASMR animations present a more stable effect in terms of EDA changes, which can consistently awaken users' emotions.

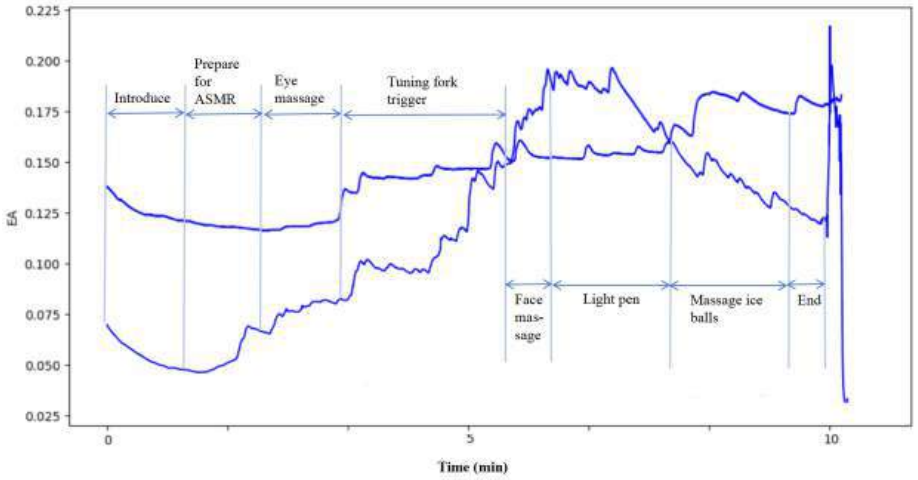


Fig. 9. EDA variability comparison.

On the other hand, the EDA data in Figure 9 shows that the two different participants have these changes at nearly the same time, which corresponds to the time of the animation settings that we have shown in Table 1. This proves that the VR-based ASMR 3D animated game that we have created successfully uses different triggers we have set up to induce ASMR sensations in users. The elevated electrodermal activity (EDA) observed reflects emotional arousal, while a decrease signifies emotional diminishment. This may be the reason why the positive scores of "happy" and "sur-prised" in the participants' emotional state scale are higher than those of convention-al VR relaxation. It can bring calmness and relaxation to users while simultaneously awakening positive emotions.

The HR data acquisition is one every 2 seconds, and after EmotiBit's self-contained data processing the HR data for each participant in an experiment is about 330. Taking 30 data per minute to calculate the average value, one participant in an experiment can get 10 heart rate averages, respectively calculate the HR average value per minute for each participant in each experiment, and finally calculate the HR average value per minute for 47 participants in different experiments can be obtained in Figure 10.

The HR of the participants decreased in all three experiments, which indicated that all three experiments were able to calm and relax the participants. Compared to the other two experiments, the VR-based ASMR animation game

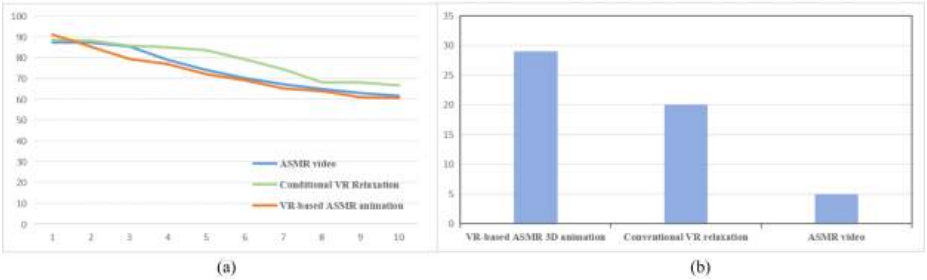


Fig. 10. (a) Comparison of the HR average value by different experiencesDA variability comparison, (b) Comparison of the number of eda data changes generated by different experiences.

showed a faster HR drop in the same amount of time, having a faster relaxation efficiency.

Since individual skin conditions vary and EDA data is affected by a variety of factors (temperature, sweat glands, body hair, etc.), it is difficult to collect the same EDA data across different participants. We assessed whether strong and continuous EDA changes occurred within a short period of time (Figures 9 and 10). As shown in Figure 12, all 47 participants experienced the three different relaxation methods. Of these, 29 exhibited strong EDA changes during the VR-based ASMR animation experience, 5 during conventional VR relaxation, and 20 while watching the ASMR video on a smartphone. These results suggest that compared to conventional VR relaxa-tion, the ASMR-related experience provides users with more intense and sustained emotional arousal. Combined with the HR reduction and the participants' emotional state change questionnaire results, we infer that VR-based ASMR provides effective relaxation in terms of sleep aid and anxiety relief while awakening positive emotions in users.

6 Results and Discussion

In this paper, we integrated VR 3D animation with ASMR to develop a VR-based ASMR animation and investigated its effectiveness in promoting relaxation and aid-ing sleep. Specifically, we detailed the production process, the development of a VR-based ASMR 3D animation game, and a comprehensive experiment designed to compare the outcomes of our method with traditional VR relaxation techniques and ASMR videos. The experimental results demonstrate that the VR-based ASMR 3D animation game can effectively alleviate negative emotions, provide positive feelings, and help reduce insomnia. Since VR-based ASMR has not been extensively studied to date, we hope our findings provide new insights for researchers interested in VR relaxation and VR avatars. In future studies we plan to incorporate MetaHuman technology and personalized characters and environments to further enhance the relaxation experience for users. Considering the variability in individual factors such as sweat glands, skin conditions, and

body hair, we will expand the participant pool to improve the accuracy of physiological data collected by the detection bracelet. Since there is a gap in research in related fields for analyzing participants' physiological data, it is difficult to find simi-lar studies for reference and comparison. In addition, there is a lack of established methodology for processing physiological data in our experiments. We will address this issue and analyze the data more comprehensively. In future research, we will also incorporate eye tracking and other monitoring methods to provide more diverse data and further strengthen the scientific support for the study.

References

1. Fredborg B, Clark J, Smith S D. An examination of personality traits associated with auton-omous sensory meridian response (ASMR)[J]. Frontiers in psychology, 2017, 8: 247. F.: Article title. Journal 2(5), 99–110 (2016)
2. Barratt E L, Spence C, Davis N J. Sensory determinants of the autonomous sensory meridi-an response (ASMR): understanding the triggers[J]. PeerJ, 2017, 5: e3846.
3. Smejka T, Wiggs L. The effects of autonomous sensory meridian response (ASMR) videos on arousal and mood in adults with and without depression and insomnia[J]. Journal of af-fective disorders, 2022, 301: 60-67.
4. Hu M Q, Li H L, Huang S Q, et al. Reduction of psychological cravings and anxiety in women compulsorily isolated for detoxification using autonomous sensory meridian re-sponse (ASMR)[J]. Brain and behavior, 2022, 12(7): e2636.
5. Navarro-Haro M V, López-del-Hoyo Y, Campos D, et al. Meditation experts try Virtual Re-ality Mindfulness: A pilot study evaluation of the feasibility and acceptability of Virtual Re-ality to facilitate mindfulness practice in people attending a Mindfulness conference[J]. PloS one, 2017, 12(11): e0187777.
6. Kaleva I, Riches S. Stepping inside the whispers and tingles: multisensory virtual reality for enhanced relaxation and wellbeing[J]. Frontiers in Digital Health, 2023, 5: 1212586.
7. Poerio G L, Blakey E, Hostler T J, et al. More than a feeling: Autonomous sensory meridian response (ASMR) is characterized by reliable changes in affect and physiology[J]. PloS one, 2018, 13(6): e0196645.
8. Sakurai N, Nagasaka K, Takahashi S, et al. Brain function effects of autonomous sensory meridian response (ASMR) video viewing[J]. Frontiers in Neuroscience, 2023, 17: 1025745.
9. Grothe-Hammer M. Tingles and society: the emotional experience of ASMR as a social phenomenon[J]. Sociological Inquiry, 2024.
10. Zappavigna M. Digital intimacy and ambient embodied copresence in YouTube videos: con-struing visual and aural perspective in ASMR role play videos[J]. Visual Communication, 2023, 22(2): 297- 321.
11. Lochte B C, Guillory S A, Richard C A H, et al. An fMRI investigation of the neural corre-lates underlying the autonomous sensory meridian response (ASMR)[J]. BioImpacts: BI, 2018, 8(4): 295.
12. Andersen J. Now you've got the shiveries: Affect, intimacy, and the ASMR whisper com-munity[J]. Television New Media, 2015, 16(8): 683-700.
13. Smith N, Snider A M. ASMR, affect and digitally-mediated intimacy[J]. Emotion, Space and Society, 2019, 30: 41- 48.

14. de Zambotti M, Barresi G, Colrain I M, et al. When sleep goes virtual: the potential of using virtual reality at bedtime to facilitate sleep[J]. Sleep, 2020, 43(12): zsaa178.
15. Schröder D, Wrona K J, Müller F, et al. Impact of virtual reality applications in the treatment of anxiety disorders: A systematic review and meta-analysis of randomized-controlled tri-als[J]. Journal of behavior therapy and experimental psychiatry, 2023, 81: 101893.
16. Goudman L, Jansen J, Billot M, et al. Virtual reality applications in chronic pain manage-ment: systematic review and metaanalysis[J]. JMIR Serious Games, 2022, 10(2): e34402.
17. Baghaei N, Chitale V, Hlasnik A, et al. Virtual reality for supporting the treatment of depres-sion and anxiety: scoping review[J]. JMIR mental health, 2021, 8(9): e29681.
18. Veling W, Lestestuiver B, Jongma M, et al. Virtual reality relaxation for patients with a psy-chiatric disorder: crossover randomized controlled trial[J]. Journal of medical Internet re-search, 2021, 23(1): e17233.
19. Pardini S, Gabrielli S, Dianti M, et al. The role of personalization in the user experience, preferences and engagement with virtual reality environments for relaxation[J]. International Journal of Environmental Research and Public Health, 2022, 19(12): 7237.
20. Cieślik B, Juszko K, Kiper P, et al. Immersive virtual reality as support for the mental health of elderly women: a randomized controlled trial[J]. Virtual Reality, 2023, 27(3): 2227-2235.
21. She Y, Wang Q, Liu F, et al. An interaction design model for virtual reality mindfulness meditation using imagery-based transformation and positive feedback[J]. Computer Anima-tion and Virtual Worlds, 2023, 34(3-4): e2184.
22. Barton A C, Do M, Sheen J, et al. The restorative and state enhancing potential of abstract fractal-like imagery and interactive mindfulness interventions in virtual reality[J]. Virtual Re-ality, 2024, 28(1): 53.
23. Aleksandrovich A, Gomes L M. Shared multisensory sexual arousal in virtual reality (VR) environments[J]. Paladyn, Journal of Behavioral Robotics, 2020, 11(1): 379-389.
24. Chung S M, Chen Z Y, Wu C T. Synaesthesia sound design in virtual reality[C]//International Conference on ArtsIT, Interactivity and Game Creation. Cham: Springer Nature Switzerland, 2022: 535-541.
25. Peng D, Pai Y S, Minamizawa K. asmVR: Light Triggers in Virtual Reality to Induce ASMR[C]//ICAT-EGVE (Posters and Demos). 2022: 17-18.
26. Peng D, Person T, Skierś K, et al. asmVR: enhancing ASMR tingles with multimodal trig-gers based on virtual reality[M]//SIGGRAPH Asia 2023 XR. 2023: 1-2.
27. Peng D, Person T, Shen X, et al. Impact of Vibrotactile triggers on mental wellbeing through ASMR experience in VR[C]//International Conference on Human Haptic Sensing and Touch Enabled Computer Applications. Cham: Springer Nature Switzerland, 2024: 398-410.
28. Poerio G L, Ueda M, Kondo H M. Similar but different: High prevalence of synesthesia in autonomous sensory meridian response (ASMR)[J]. Frontiers in Psychology, 2022, 13: 990565.
29. Niu S, Manon H S, Bartolome A, et al. Close-up and whispering: an understanding of mul-timodal and parasocial interactions in YouTube ASMR videos[C]//Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022: 1-18.

30. Barratt E L, Davis N J. Autonomous Sensory Meridian Response (ASMR): a flow-like mental state[J]. PeerJ, 2015, 3: e851.
31. Montgomery S M, Nair N, Chen P, et al. Introducing EmotiBit, an open-source multi-modal sensor for measuring research-grade physiological signals[J]. Science Talks, 2023, 6: 100181.
32. Critchley H D. Electrodermal responses: what happens in the brain[J]. The Neuroscientist, 2002, 8(2): 132-142.
33. Engelbregt H J, Brinkman K, Van Geest C C E, et al. The effects of autonomous sensory meridian response (ASMR) on mood, attention, heart rate, skin conductance and EEG in healthy young adults[J]. Experimental Brain Research, 2022, 240(6): 1727-1742.

Automating Visual Narratives: Learning Cinematic Camera Perspectives from 3D Human Interaction

Boyuan Cheng[1], Shang Ni[1], Jian Jun Zhang[1], and Xiaosong Yang[1]

National Centre for Computer Animation, Bournemouth University, Fern Barrow, Poole, Dorset, BH12 5BB, United Kingdom
{bcheng,s5701147,jzhang,xyang}@bournemouth.ac.uk

Abstract. Cinematic camera control is a cornerstone of visual storytelling in film, animation, and interactive media, yet remains a labor-intensive task typically handled by expert artists. While recent deep learning methods automate camera placement and movement from video, they depend heavily on large, annotated video corpora and struggle to generalize to novel character interactions. In this work, we propose a novel framework that learns to predict Toric camera parameters directly from two-person 3D motion data, bypassing the need for preexisting visual datasets. Our model employs a dual-stream Transformer to encode each character's motion, fuses these streams via bidirectional cross-attention to capture inter-character dynamics, and incorporates explicit spatial vectors to ground geometric relationships. A lightweight fusion network then regresses per-frame Toric parameters, yielding smooth, compositionally balanced camera trajectories. To enable training and evaluation, we introduce a new dataset of over 3,400 motion–camera sequences spanning diverse interaction scenarios. Experiments demonstrate that our approach significantly outperforms a strong Example-Driven Camera baseline and ablated variants in trajectory accuracy, framing quality, and temporal coherence.

Keywords: Virtual cinematography · 3D human motion · Computer animation.

1 Introduction

Animation is widely recognized as a compelling storytelling medium, uniquely capable of delivering narratives through visual framing. Expressive character portrayals, richly detailed environments, and nuanced lighting collectively form the foundation of animated storytelling. However, the role of cinematography, particularly shot composition and camera positioning, in influencing audience emotion and narrative coherence is frequently underappreciated in the animation industry. Effective shot design goes beyond selecting camera angles or movement tra-

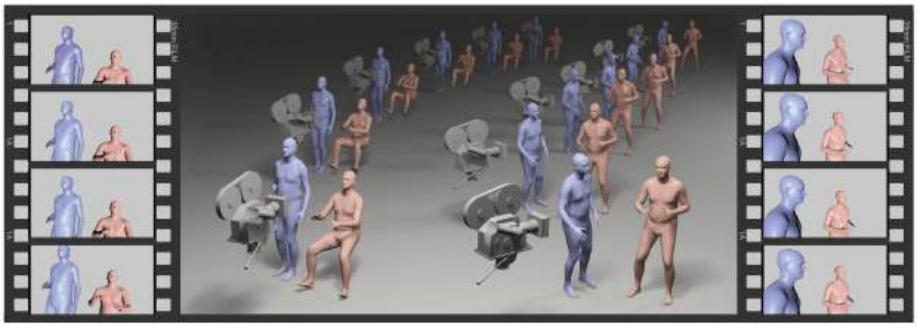


Fig. 1. Our method automatically generates cinematic camera motions from 3D interactive motion sequence (left), producing smooth, compositionally balanced camera views (right) that follow professional filmmaking conventions.

jectories; it requires intentional planning and precise positioning to clearly communicate character interactions, relationships, and psychological states. Therefore, the camera serves as an essential tool enabling directors to articulate artistic intentions, significantly enhancing narrative clarity and emotional resonance.

Despite its critical role, identifying optimal cinematographic compositions in animation is traditionally a complex and labor-intensive task, demanding significant expertise from directors and specialized artists. Typically, production teams rely on manual refinement and iterative experimentation to finalize camera movements and settings. This conventional approach is inefficient, often escalating production costs and creating technical obstacles. Small and medium-sized animation studios are particularly affected, as they often lack access to skilled personnel and sufficient technical resources. Consequently, these limitations hinder creative exploration, negatively impacting the overall quality and expressive potential of animated narratives.

Recent advancements in deep learning offer promising avenues to automate the processes of camera blocking and shot composition. Current deep learning methods successfully estimate camera motion and framing from existing visual datasets, closely replicating professional cinematographic techniques. However, a critical drawback of these techniques is their dependency on extensive, predefined visual datasets. This dependence makes their performance highly sensitive to the diversity, quality, and representative nature of the available training data, limiting their adaptability and effectiveness when dealing with novel character interactions or unfamiliar animation contexts.

To address these limitations, this paper introduces a novel method that eliminates reliance on predefined visual datasets by leveraging 3D motion data to estimate camera placements and cinematographic composition (Figure 1). Our approach specifically processes interactions between pairs of animated characters, utilizing deep learning integrated with Toric features to capture spatial orientations and relative positions in 3D space.

In summary, this paper makes the following key contributions:

- To the best of our knowledge, this is the first work to explore how 3D human motion can be leveraged to learn cinematic camera movement, bridging the gap between character motion and cinematographic principles.
- We construct a dataset that explicitly links human motion with camera parameters, providing a valuable resource for future research in motion-aware camera control.
- We propose a novel framework that models the spatio-temporal dynamics between character motion and camera movement, enabling more natural and expressive cinematography.
- Our model outperforms baseline methods and ablation variants, demonstrating its effectiveness in learning cinematic camera motion directly from character movement.

2 Related Works

Designing dynamic camera movements [1–7] poses a formidable challenge, shaped by multiple factors and spurring considerable inquiry into methods of synthesis and control. Early work regarded the camera planning task as a constraint-satisfaction problem, using constraint-based optimization to achieve desired camera behaviors [8–10]. With the rise of deep learning, neural network-driven approaches have grown increasingly prevalent. Jiang et al. assembled a film clip dataset, which encompasses camera movement and actor motion, to investigate synthesis from film references or textual prompts via LSTM and diffusion models [11–14]. Meanwhile, Wu et al. introduced a GAN-based controller designed to produce camera movements suited to storytelling contexts [4]. Additional progress includes techniques for transferring cinematic effects to 3D virtual environments, one of the studies develops a differentiable pipeline to estimate and optimize camera and character motions from existing films, facilitating retargeting to 3D engines [15].

In gaming, considerable effort has been invested in automated camera control to elevate player engagement. Li and Cheng introduced a module for third-person tracking, while Rucks and Katzakis [5] created CameraAI to reduce occlusions during pursuit phases. Evin et al. [16] further incorporated recognized cinematographic standards into a semi-automated system, Cine-AI, to produce immersive in-game cutscenes.

Automatically generating camera work for dance sequences poses additional hurdles, given the interplay of shot variety, music, and intricate dance movements. Xie et al. [17] explored deriving camera trajectories from dancer poses, though their solution did not fully integrate musical influences and required extra keyframe data. To address these issues, Wang et al. [1] presented the first 3D dance-camera-music dataset (DCM) alongside a transformer-based diffusion model, DanceCamera3D, to tackle this challenge. However, their approach relies on smoothing to handle abrupt transitions, which can detract from the effectiveness of sudden camera switches.

Other research has attempted to compute camera parameters between keyframes using neural networks [17, 14], yet these techniques often result in jitter or otherwise unsatisfactory motion, prompting additional smoothing or resorting to simpler 2D formats that can curtail creative flexibility. Consequently, refining camera movement—particularly within 3D dance contexts—remains an ongoing and demanding area of investigation.

3 Dataset

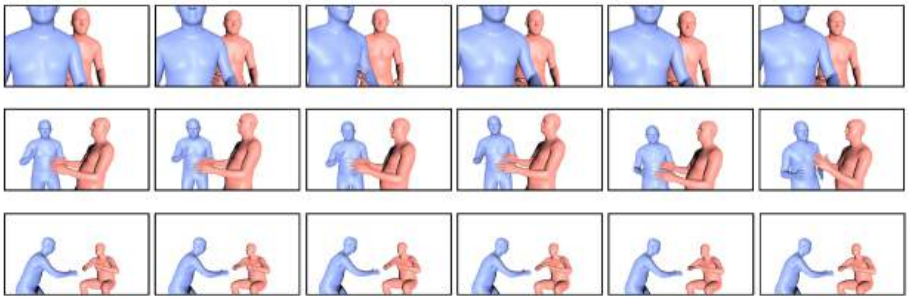


Fig. 2. Sampled motion-camera sequences from our dataset. Each row shows a 6-second human interaction clip with corresponding dynamic camera views.

We construct a motions-camera dataset specifically tailored for the motion-guided cinematography generation task. Our dataset comprises 3442 motions-camera sequences, each capturing dynamic interactions and camera responses. Every sequence spans 6 seconds, corresponding to 120 frames captured at 20 frames per second. These sequences encompass diverse interaction patterns, spatial relationships, and dynamic variations between characters, along with corresponding professional-level camera movements. This dataset facilitates the development and training of deep-learning models capable of generating expressive, coherent, and cinematic camera movements guided purely by dual-character 3D motion data.

We extract motion sequences directly from video clips, which are in turn sourced from raw full-length films, television series, and stage performances. The selection of these clips leverages annotations from SHOTDECK, a comprehensive online database of cinematic shots, which provides detailed tagging information such as character count, camera angles, and exact timestamps. Each retained SHOTDECK shot is then mapped back to its original video source using the provided timestamp as a reference point. Around these timestamps, we manually inspect and extract continuous segments from the original videos, focusing exclusively on clips lasting at least 6 seconds without any intervening camera transitions.

For each selected video clip, we first detect and track the bounding boxes of the two main characters frame-by-frame. Bounding boxes are propagated across consecutive frames by calculating the Intersection-over-Union (IoU), ensuring consistent character identification and reliable tracking throughout the entire sequence. Subsequently, we apply MeTRAbs [18], which estimates 2D keypoints and relative 3D pose and applies perspective geometry optimization to recover the absolute 3D root position. Each character's pose is represented using the SMPL-22 joint model [19]. Finally, to enhance temporal smoothness and motion coherence, a Gaussian filter is applied to each joint trajectory along the temporal dimension, resulting in a set of two-character 3D motion sequences. In MeTRAbs, the camera's 3D coordinates are defined at the origin, however, we do not directly use these raw camera coordinates as camera features. The specific feature transformations applied for camera representation will be detailed in section 4.2. Figure 2 provides visual examples of our dataset, illustrating extracted motion data alongside camera data.

4 Methodology

4.1 Motion Data Representation

The motion of a single character over N frames is expressed as $x^{1:N} = \{x^i\}_{i=1}^N$, where each x^i encodes the pose information at frame i. Specifically, x^i consists of the rotations of 22 out of the 24 joints in the SMPL model (excluding the two hand joints) and the 3D global translation of the root, which together define a rigged and skinned 3D character. Each joint rotation is represented by a 3×3 matrix, but instead of using the full matrix, only its first two orthogonal column vectors are retained, resulting in a 6D representation per joint. These joint rotations are estimated through inverse kinematics from the corresponding 3D joint positions. To represent root translation in a compatible form, it is converted into a 6D vector by appending three zeros before concatenating it with the rotation vectors. Consequently, x^i is stored as a 23×6 matrix, which is then flattened into a 138-dimensional vector. The full motion sequence is thus represented as a tensor of shape $N \times 138$.

For scenarios involving two characters, their motion representations are combined frame-wise. However, since each character's motion is originally represented in its own local coordinate system where the root is fixed at the origin in the first frame, directly merging their sequences into a shared 3D space can lead to significant overlaps between the two characters over time. To prevent this, an offset vector $D \in R^9$ is introduced to encode each character's initial orientation and position relative to the global coordinate system at the first frame. The first six elements of D capture the character's orientation, represented by the first two orthogonal vectors of its rotation matrix, while the last three elements define its initial position in the global frame. The character's facing direction is determined by assuming that the 3D line connecting its shoulder joints remains parallel to the xz-plane, with the angle between this line and the x-axis defining

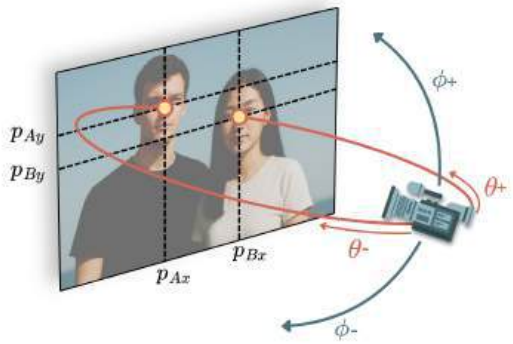


Fig. 3. Toric parametrization of a two-shot. The camera can *orbit* around the baseline AB by an azimuth φ (blue arrows) and *raise* or *lower* itself with an elevation θ (red arrows) while guaranteeing that both characters remain on screen.

its initial orientation. Notably, D is computed solely from the first frame but is applied consistently across all subsequent frames.

As a result, a two-character motion sequence is represented as $(x_A^{1:N}, x_B^{1:N}, D_A, D_B)$, where $(x_A^{1:N}, x_B^{1:N})$ denote the motion representations of the two characters in their respective local coordinate frames, and (D_A, D_B) encode their initial spatial offsets in the global frame.

4.2 Camera Data Representation

This project follows the Toric formulation of Lino and Christie [7], because it couples camera motion to the on-screen layout of two protagonists A and B while remaining independent of focal length and aspect ratio (Fig. 3). To compute the Toric coordinates, we utilize data extracted from the raw videos, which provides the camera position in world coordinates x_C^i (considered as the origin), the world coordinates of the character x_A^i, x_B^i, and their corresponding screen-space coordinates p_A^i, p_B^i.

Opening angle: Since the field of view (FOV) f and aspect ratio l are known, the screen-space coordinates are transformed into normalized device coordinates (NDC), allowing the computation of sight-line opening $\alpha^i \in (0, \pi)$ directly from screen-space information:

$$S_x = \frac{1}{\tan\left(\frac{f}{2}\right)} \tag{1}$$

$$S_y = S_x\, l \tag{2}$$

$$p_k'^i = \left(\frac{p_{kx}^i}{S_x},\ \frac{p_{ky}^i}{S_y},\ 1\right), \quad k \in \{A, B\} \tag{3}$$

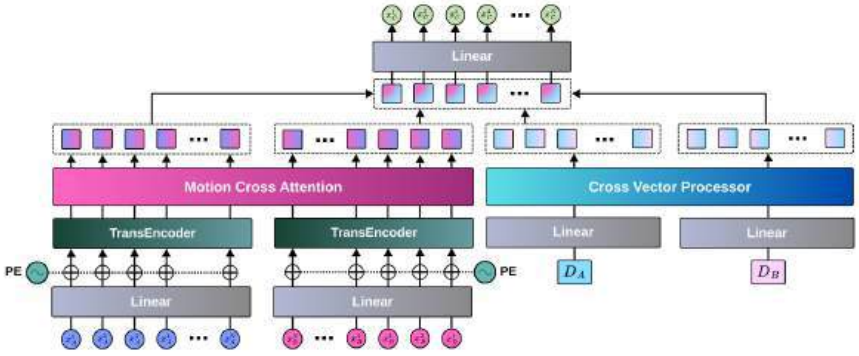


Fig. 4. Overview of our dual-stream Toric camera prediction architecture. Per-frame joint positions for persons A and B ($x_A^{1:N}, x_B^{1:N}$) are first linearly projected and augmented with learnable positional encodings, then processed by separate Transformer encoder stacks to produce sequence features. These two streams interact via a Multi-Head Motion Cross-Attention module, which allows each person's motion context to inform the other. Simultaneously, the per-frame relative spatial vectors D_A and D_B are mapped through lightweight linear layers and fused by the Cross Vector Processor into geometric feature sequences. The four resulting feature streams are concatenated and passed through a final linear layer to regress the per-frame Toric camera parameters $x_C^{1:N}$.

$$\alpha^i = \arccos\left(\frac{p'^i_A \cdot p'^i_B}{||p'^i_A|| \cdot ||p'^i_B||}\right) \tag{4}$$

where p'^i_A and p'^i_B are the normalized screen-space projections of the character.

In-plane basis: Let u^i be the unit vector from target A to target B, and let $w = (0, 1, 0)^T$ denote the world-up direction. Project w onto the plane orthogonal to u^i,

$$r = normalize\big(w - (w \cdot u^i)u^i\big) \tag{5}$$

and obtain a second in-plane axis by a 90° rotation,

$$t^i = u^i \times r^i \tag{6}$$

The pair (r^i, t^i) thus forms an orthonormal basis of the plane perpendicular to the baseline AB.

Elevation and azimuth: Denote the midpoint of the baseline by m^i and the camera offset by $v^i = x_C^i - m^i$. The elevation $\theta^i \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ and the azimuth $\varphi^i \in (-\pi, \pi]$ are recovered from the offset vector v^i using the following expressions:

$$\theta^i = atan2\big(v^i \cdot u^i, v^i \times u^i\big) \tag{7}$$

$$\varphi^i = atan2\big(v^i \cdot t^i, v^i \cdot r^i\big) \tag{8}$$

Learning descriptor: Since α is uniquely determined by these two points, our per-frame descriptor stores only the screen positions and the two control angles

$$c^{1:N} = \{p_A^i, p_B^i, \theta^i, \phi^i\}_{i=1}^N \in R^{6N} \tag{9}$$

This 6D sequence tightly couples the camera trajectory to the actors' screen composition, guarantees they remain in shot, and provides a compact manifold on which our network learns cinematographically plausible camera motion.

4.3 Network Structure

We propose a novel neural architecture to predict camera control parameters. Given two motion sequences and their initial spatial configuration, our model integrates dual-stream temporal encoding with structured spatial reasoning to infer how the camera should behave in response to the evolving motion context. Each character's motion sequence is first projected into a latent space and encoded via a multi-layer Transformer encoder with learnable positional encoding. Formally, for motion inputs $x_j^{1:N} \in R^{N\times 138}$ with $j \in \{A, B\}$, the temporal features are obtained as:

$$z_j^{1:N} = x_j^{1:N} W_e + P \tag{10}$$

$$h_j^{1:N} = \text{softmax}\left(\frac{z_j^{1:N} W_Q (z_j^{1:N} W_K)^T}{\sqrt{E}}\right) z_j^{1:N} W_V \tag{11}$$

where W_e is the shared embedding matrix, P is a learnable positional encoding, and W_Q, W_K, W_V are the standard self-attention projections. This time series Transformer module encodes each character's motion in context, enabling the model to reason about movement evolution over time.

To explicitly model inter-character interaction, we introduce a Motion Cross-Attention module that performs bidirectional attention between the two encoded motion sequences. Specifically, each character's representation is refined by attending to the other's contextual features:

$$\tilde{h}_A^{1:N} = h_A^{1:N} + \text{softmax}\left(\frac{h_A^{1:N} W_Q (h_B^{1:N} W_K)^\top}{\sqrt{E}}\right) h_B^{1:N} W_V \tag{12}$$

$$\tilde{h}_B^{1:N} = h_B^{1:N} + \text{softmax}\left(\frac{h_B^{1:N} W_Q (h_A^{1:N} W_K)^\top}{\sqrt{E}}\right) h_A^{1:N} W_V \tag{13}$$

This mechanism refines with cues from and vice-versa, enabling the network to model reaction timing, synchronised gestures and antagonistic behaviours that are invisible to single-stream encoders.

Static spatial cues are injected through two weight-sharing Cross Vector Processors. Each 9-D global pose vector is linearly projected and cross-attended against the partner's projection, producing relation-aware embeddings. These embeddings are broadcast along the temporal axis so every frame is conditioned on the initial spatial relationship.

The temporally contextualized motion features (from both characters) and their respective spatial embeddings are concatenated along the feature dimension. This representation is processed through a fully connected fusion network composed of linear transformations, ReLU activations, and layer normalization, and finally projected to an output of dimension, which corresponds to the predicted Toric parameters at each time step.

5 Experiments

We conduct our experiments using a dataset where 80% of the samples are allocated for training, while the remaining 20% are used for testing. The model is trained for 20,000 steps with a batch size of 32. We utilize an Adam optimizer with a learning rate of 1×10^{-5} to update the model parameters. To evaluate the effectiveness of our approach, we perform ablation studies to analyze the impact of different model components. Additionally, we conduct comparative experiments against baseline models. The results are quantitatively analyzed through standard evaluation metrics, and we further validate our findings via a user study, gathering qualitative feedback on the model's applicability.

5.1 Evaluation Metrics

To evaluate the generated camera motions, we adopt three complementary metrics: *FrameFID*, *SeqFID* and Pose Distance Error (*PDE*). *FrameFID* measures the framing quality of individual frames based on the ratio of visible body parts and their screen projections, following Wang et al. [1]. *SeqFID* quantifies the distributional distance between generated camera motions and real motions from the dataset. It is computed using features extracted by a self-supervised VAE-Transformer encoder. *PDE* directly assesses geometric fidelity by averaging, over all frames, the positional discrepancy between predicted and ground-truth camera features.

5.2 Comparison to Baselines

We compare our model with baseline method of Jiang et al. [13]. To ensure a fair comparison, we process our two-person motion sequences to exactly match Jiang et al.'s input requirements: we keep the same frame rate and skeleton topology, convert our keypoint data into the official ActionGraph encoding per-frame world positions, orientation vectors, and velocity information for both characters, and then feed this directly into their proposed Mixture-of-Experts (MoE) network. The quantitative results in Table 1 demonstrate that our method substantially outperforms the Jiang et al. baseline across all three evaluation metrics. We observe a significant reduction in Pose Distance Error, indicating more accurate recovery of the camera's spatial trajectory. Likewise, our FrameFID is dramatically lower, reflecting greatly improved per-frame composition with consistently well-centered subjects and minimal clipping. Finally, the considerable drop in SeqFID confirms that our predicted camera motions are far smoother and more temporally coherent than those produced by the baseline. Figure 5 provides qualitative support for these findings by juxtaposing ground-truth frames (top row), Jiang et al.'s outputs (middle row), and our predictions (bottom row) across four representative interaction sequences. While the baseline often exhibits abrupt viewpoint shifts, frame-to-frame jitter, and occasional loss of one subject from the viewport, our method maintains stable, cinema-style framing and fluid camera trajectories that closely follow the ground truth, even during rapid movements and partial occlusions.

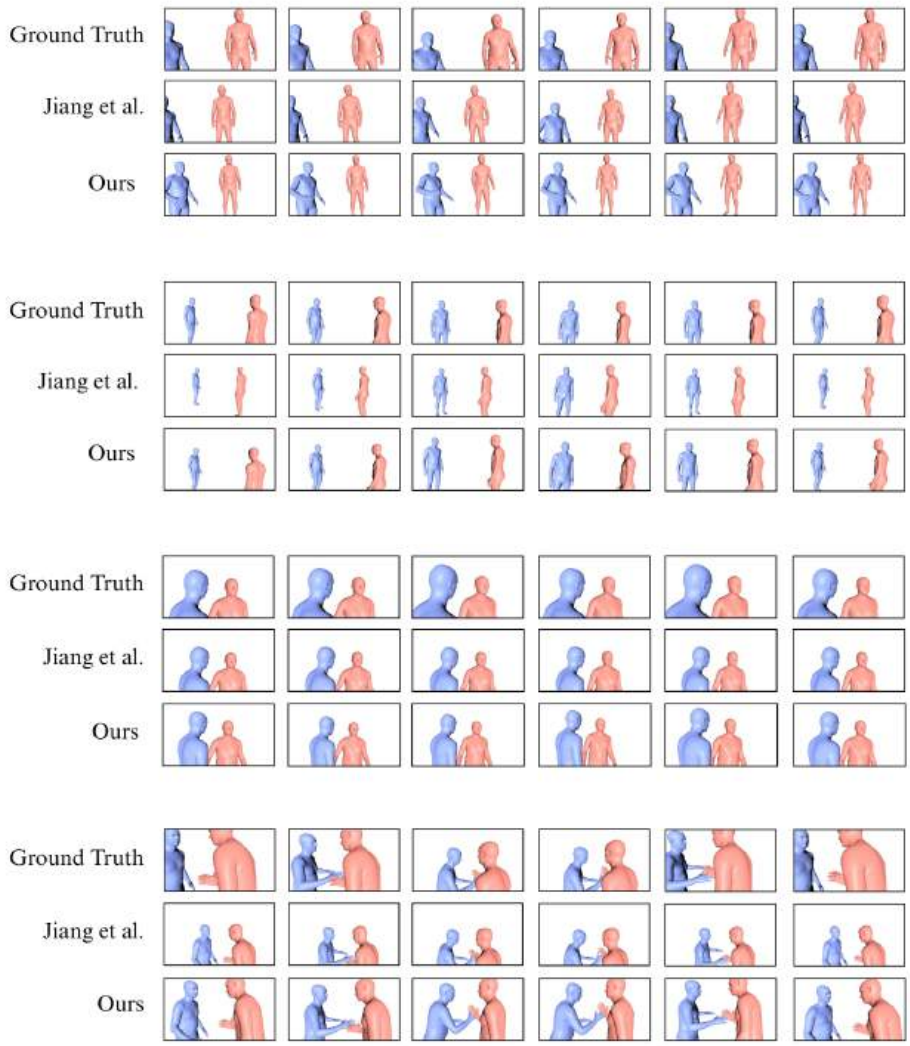


Fig. 5. Qualitative comparison on four representative two-person interaction sequences. Each column shows a temporal snapshot, with rows depicting (top) the ground-truth camera view, (middle) the Example-Driven Camera baseline (Jiang et al.), and (bottom) our method. While the baseline often produces abrupt viewpoint shifts and framing inconsistencies under rapid movements, our approach maintains smooth, stable composition that closely follows the ground truth throughout each interaction.

Table 1. Quantitative comparison of our model against the model proposed by Jiang et al. Both are trained on our dataset.

Method	*PDE* ↓	*FrameFID* ↓	*SeqFID* ↓
Jiang et al.	0.475	2.549	2.997
Ours	**0.451**	**0.443**	**0.628**

Table 2. Results of the ablation study, where one component is removed at one time to test its importance.

Method	*PDE* ↓	*FrameFID* ↓	*SeqFID* ↓
w/o Offset	0.644	1.128	2.377
w/o Att	0.473	0.443	0.850
Ours	**0.451**	**0.368**	**0.628**

5.3 Ablation Study

To validate the individual contributions of our cross-attention module and explicit spatial vector input, we conducted an ablation study with two variants: one without cross-attention and one without spatial vectors. Table **??** reports the ablation results. When the spatial offset input is removed (w/o Offset), all three metrics degrade most severely: the *PDE* rises sharply and both *FrameFID* and *SeqFID* worsen substantially, underscoring the critical role of explicit geometric cues in maintaining framing accuracy and temporal coherence. Omitting the cross-attention module (w/o Att) leads to a moderate drop in all metrics, demonstrating its importance for modeling interactions between characters. By contrast, our full model, which integrates both spatial offsets and cross-attention, achieves the lowest *PDE*, *FrameFID*, and *SeqFID*, demonstrating that these components work synergistically to produce precise, smooth, and cinema-style camera motions.

6 Conclusion

In this paper, we have introduced a novel framework for predicting cinematic camera motions directly from dual-character 3D motion data, without relying on large pre-existing video corpora. Central to our approach is the integration of cross-attention for modeling inter-character interactions and explicit spatial vectors for encoding global geometry, all within a dual-stream Transformer architecture. We also contribute a new motions-camera dataset that tightly couples professional camera trajectories with dynamic two-person interactions. Quantitative evaluations against a strong Example-Driven Camera baseline and extensive ablations demonstrate that our method produces more accurate, stable, and visually coherent Toric camera trajectories. In future work, we plan to extend this approach to multi-character scenarios, incorporate stylistic controls for finer

cinematographic effects, and explore real-time deployment in interactive virtual environments.

References

1. Z. Wang, J. Jia, S. Sun, H. Wu, R. Han, Z. Li, D. Tang, J. Zhou, and J. Luo, "DanceCamera3D: 3D camera movement synthesis with music and dance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7892–7901, 2024.
2. R. Courant, N. Dufour, X. Wang, M. Christie, and V. Kalogeiton, "ET the Exceptional Trajectories: Text-to-Camera-Trajectory Generation with Character Awareness," in *European Conference on Computer Vision*, pp. 464–480, Springer, 2024.
3. A. Rao, X. Jiang, Y. Guo, L. Xu, L. Yang, L. Jin, D. Lin, and B. Dai, "Dynamic Storyboard Generation in an Engine-Based Virtual Environment for Video Production," in *ACM SIGGRAPH 2023 Posters*, pp. 1–2, 2023.
4. X. Wu, H. Wang, and A. K. Katsaggelos, "The Secret of Immersion: Actor Driven Camera Movement Generation for Auto-Cinematography," *arXiv preprint arXiv:2303*, 2023.
5. J. Rucks and N. Katzakis, "CameraAI: Chase Camera in a Dense Environment Using a Proximal Policy Optimization-Trained Neural Network," in *2021 IEEE Conference on Games (CoG)*, pp. 1–8, IEEE, 2021.
6. Z. Wang, J. Li, X. Qin, S. Sun, S. Zhou, J. Jia, and J. Luo, "DanceCamAnimator: Keyframe-Based Controllable 3D Dance Camera Synthesis," in *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10200–10209, 2024.
7. C. Lino and M. Christie, "Intuitive and Efficient Camera Control with the Toric Space," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 4, pp. 1–12, 2015.
8. M. Christie, P. Olivier, and J.-M. Normand, "Camera Control in Computer Graphics," in *Computer Graphics Forum*, vol. 27, pp. 2197–2218, Wiley Online Library, 2008.
9. W. Bares, S. McDermott, C. Boudreaux, and S. Thainimit, "Virtual 3D Camera Composition from Frame Constraints," in *Proceedings of the Eighth ACM International Conference on Multimedia*, pp. 177–186, 2000.
10. M. Christie and J.-M. Normand, "A Semantic Space Partitioning Approach to Virtual Camera Composition," in *Computer Graphics Forum*, vol. 24, pp. 247–256, 2005.
11. J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.
12. H. Jiang, M. Christie, X. Wang, L. Liu, B. Wang, and B. Chen, "Camera Keyframing with Style and Control," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 6, pp. 1–13, 2021.
13. H. Jiang, B. Wang, X. Wang, M. Christie, and B. Chen, "Example-Driven Virtual Cinematography by Learning Camera Behaviors," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, p. 45, 2020.
14. H. Jiang, X. Wang, M. Christie, L. Liu, and B. Chen, "Cinematographic Camera Diffusion Model," in *Computer Graphics Forum*, vol. 43, p. e15055, Wiley Online Library, 2024.
15. X. Jiang, A. Rao, J. Wang, D. Lin, and B. Dai, "Cinematic Behavior Transfer via NeRF-Based Differentiable Filming," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6723–6732, 2024.

16. I. Evin, P. Hämäläinen, and C. Guckelsberger, "Cine-AI: Generating Video Game Cutscenes in the Style of Human Directors," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6(CHI PLAY), pp. 1–23, 2022.
17. C. Xie, I. Hemmi, H. Shishido, and I. Kitahara, "Camera Motion Generation Method Based on Performer's Position for Performance Filming," in *2023 IEEE 12th Global Conference on Consumer Electronics (GCCE)*, pp. 957–960, IEEE, 2023.
18. I. Sárándi, T. Linder, K. O. Arras, and B. Leibe, "MeTRAbs: Metric-Scale Truncation-Robust Heatmaps for Absolute 3D Human Pose Estimation," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 1, pp. 16–30, 2021.
19. M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A Skinned Multi-Person Linear Model," *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

Peridynamics-Based Simulation of Viscoelastic Solids and Granular Materials

Jiamin Wang[1,2][0009−0006−1151−5802], Haoping Wang[1], Xiaokun Wang[1], Yalan Zhang[1], Jiří Kosinka[2][0000−0002−8859−2586], Steffen Frey[2][0000−0002−1872−6905], Alexandru Telea[3], and Xiaojuan Ban[1]

[1] School of Intelligence Science and Technology, University of Science and Technology Beijing, Beijing, China
[2] Bernoulli Institute, University of Groningen, Groningen, the Netherlands
jiamin.wang@rug.nl
Faculty of Science, Utrecht University, Utrecht, the Netherlands

Abstract. Viscoelastic solids and granular materials have been extensively studied in Classical Continuum Mechanics (CCM). However, CCM faces inherent limitations when dealing with discontinuity problems. Peridynamics, as a non-local continuum theory, provides a novel approach for simulating complex material behavior. We propose a unified viscoelastoplastic simulation framework based on State-Based Peridynamics (SBPD) which derives a time-dependent unified force density expression through the introduction of the Prony model. Within SBPD, we integrate various yield criteria and mapping strategies to support granular flow simulation, and dynamically adjust material stiffness according to local density. Additionally, we construct a multi-material coupling system incorporating viscoelastic materials, granular flows, and rigid bodies, enhancing computational stability while expanding the diversity of simulation scenarios. Experiments show that our method can effectively simulate relaxation, creep, and hysteresis behaviors of viscoelastic solids, as well as flow and accumulation phenomena of granular materials, all of which are very challenging to simulate with earlier methods. Furthermore, our method allows flexible parameter adjustment to meet various simulation requirements.

Keywords: Peridynamics · Viscoelastic simulation · Granular materials · Multi-material coupling.

1 Introduction

Viscoelastic solids and granular materials are ubiquitous in our daily lives and industrial production. From kneading dough and biological soft tissues to natural disasters like avalanches and mudflows, these materials demonstrate complex dynamic characteristics. Accurate simulation of these behaviors is of great significance to fields such as materials science, geotechnical engineering, biomedical simulation, and – last but not least – computer graphics.

Viscoelastic solids have time-dependent characteristics including *stress relaxation*, *creep*, and *hysteresis*. For large deformations, memory effects and nonlinearities further complicate the simulation. Granular materials consist of a

large number of discrete particles and can exhibit both the shear resistance of solids and the deformability of fluids. Recent advances in computational and physical modeling techniques have made the accurate simulation of viscoelastic and granular materials an active area of research in both computer graphics and computational physics.

Early simulation methods used mesh-based discretization strategies such as the Finite Element Method (FEM) [25] and described the time-dependent behavior of viscoelastic materials by generalized Maxwell or Kelvin-Voigt models. While widely used in structural mechanics, when handling fractures, separations, and large deformations, such models encounter complex challenges when topology changes and meshes need reconstruction. Mesh-free methods such as Smoothed Particle Hydrodynamics (SPH), the Material Point Method (MPM), and Position-Based Dynamics (PBD) compute physical interactions through particle-based interactions and show clear advantages in handling fracture, large deformation, and free surface flows. They can model a wide range of natural phenomena and materials such as muscle [15], sand [7, 27], snow [19, 5], and multi-material mixtures [21, 4].

However, most existing mesh-free methods still rely on CCM with foundations in partial differential equations (PDEs). PDEs are not applicable at *discontinuities*, *e.g.*, cracks and interface slippage; additional techniques are needed to capture such phenomena. The Peridynamics method [16] replaces differential with integral equations to naturally handle material discontinuities. State-Based Peridynamics (SBPD) [17] further expanded the range of constitutive models by introducing the *deformation state* and *force state* concepts. While some viscoelastic and elastoplastic models have been developed within the Peridynamics framework, the potential for granular flow simulation and unified coupling with elastic bodies remains underexplored.

In this paper, we propose a unified viscoelasto-plastic simulation framework based on SPBD that supports both the time-dependent behavior of viscoelastic solids and the yield-driven flow dynamics of granular materials, with the following key contributions:

- We introduce the Prony model to an SPBD-bsed framework to derive time-dependent force density expressions, accurately capturing *relaxation*, *creep*, and *hysteresis*.

- We integrate various yield criteria and plastic mapping strategies within SBPD, combine them with dynamic and static friction forces and density-based stiffness adjustments, and achieve realistic granular flows.

- We create a multi-material coupling system supporting *interactions* between viscoelastic solids, granular materials, and rigid bodies. This improves computational stability and significantly enriches the diversity of simulation scenarios.

2 Related Work

2.1 Viscoelastic Simulation

Viscoelastic materials under external loads exhibit both only equilibrium elastic responses and non-equilibrium viscous characteristics.

Terzopoulos and Fleischer [23, 22] pioneered the use of elastic models into computer graphics and expanded them into three typical non-elastic behavior simulations including viscoelasticity, plasticity, and fracture. Müller et al. [11] introduced SPH into computer graphics, greatly promoting the application of meshless methods in deformable body simulation. Takahashi et al. [20] proposed an implicit SPH method for stable simulation of highly viscous fluids. Peer et al. [14], by extracting rotation from the SPH deformation gradient, improved the efficiency of elastic solid simulation nearly hundredfold. The MPM [19] is a particle-grid hybrid method initially introduced to graphics primarily for snow simulation, and subsequently extended to handle many materials and phase transitions [24]. Yue et al. [26] used MPM to simulate shear-dependent dense foams. Current research on viscoelasticity in computer graphics primarily focuses on viscoelastic fluids and much less on viscoelastic solids. Fang et al. [3] proposed a predictor-corrector algorithm that achieves viscoelastic and elastoplastic solid simulation under large deformation conditions.

Peridynamics has attracted increasing interest due to its advantages in handling material failure problems such as cutting and crack propagation [1]. Yet, developing systematic viscoelastic models within a peridynamics framework remains limited. Madenci et al. [10] proposed a viscoelastic constitutive model based on ordinary state-based peridynamics, capturing material relaxation characteristics under mechanical and thermal loads. Ozdemir et al. [13] further modeled crack propagation in films based on this approach. Our method differs from theirs; although also based on the Prony model, we have derived a unified force density expression by combining it with a corotational elastic energy model.

2.2 Granular Flow Simulation

Continuum methods have been widely used in graphics to simulate granular materials. Zhu and Bridson [28] simulated sand through an improved PIC fluid solver. Narain et al. [12] made key improvements to this method, effectively eliminating cohesive artifacts related to incompressibility, significantly enhancing simulation quality. Lenaerts and Dutre [9] implemented coupling interactions between water and sand based on the SPH method. Daviet and Bertails-Descoubes [2] developed a MPM-based granular material model that behaves like a solid due to internal friction, representing granular matter as a viscoplastic fluid combining the Drucker-Prager yield criterion and unilateral compressibility constraints. Tampubolon et al. [21] proposed a multi-phase MPM simulation of sand-water mixtures, handling fluid permeation and interaction in sand via porous media theory.

Compared to SPH and MPM methods, Peridynamics-based simulation of granular materials is an emerging research direction with great potential. In structural mechanics, Peridynamics is commonly used to simulate the fracture of geotechnical materials under loading [8]. However, current research on Peridynamics for simulating granular flows remains relatively limited, particularly lacking a framework that unifies viscoelastic response with granular plastic flow.

3 SBPD Theory

State-based Peridynamics (SBPD) is a reformulation of continuum mechanics. Unlike bond-based peridynamics, which models particle interactions as springs, SBPD defines interactions through the relation between a particle and its neighborhood. This allows for asymmetric forces and the modeling of more complex material behavior.

Let $\mathcal{H}$ denote a spherical neighborhood of radius r and center $\mathbf{x}_i$. Let $\mathcal{L}_m$ denote the space of order-m tensors. An order-m *state* is a mapping $\mathbf{A}\langle\boldsymbol{\xi}\rangle : \mathcal{H} \to \mathcal{L}_m$, where $\boldsymbol{\xi} = \mathbf{x}_j - \mathbf{x}_i, \boldsymbol{\xi} \in \mathcal{H}$ is the so-called *bond vector* between particle $\mathbf{x}_i$ and its neighbor $\mathbf{x}_j$. Let $\mathbf{y} = \varphi(\mathbf{x})$ denote a deformation under a motion φ. The corresponding reference and deformation vector states (see Fig. 1) are defined as $\mathbf{X}\langle\boldsymbol{\xi}\rangle = \mathbf{x}_j - \mathbf{x}_i$ and $\mathbf{Y}\langle\boldsymbol{\xi}\rangle = \mathbf{y}_j - \mathbf{y}_i$.

Classical continuum mechanics defines the deformation gradient as $\mathbf{F}(\mathbf{x}) = \partial\mathbf{y}/\partial\mathbf{x}$. Yet, this partial derivative does not exist at discontinuities. To overcome this, Peridynamics approximates $\mathbf{F}$ using a least-squares minimization over $\mathcal{H}$ as $\mathbf{F} = (\mathbf{Y} * \mathbf{X})\,(\mathbf{X} * \mathbf{X})^{-1}$ with the generalized tensor product defined by

$$\mathbf{A} * \mathbf{B} = \int_{\mathcal{H}} w(\boldsymbol{\xi})\, \mathbf{A}\langle\boldsymbol{\xi}\rangle \otimes \mathbf{B}\langle\boldsymbol{\xi}\rangle \, d\boldsymbol{\xi}, \tag{1}$$

where $w(\boldsymbol{\xi})$ is a weight function and $\otimes$ denotes the dyadic product.

The motion of particle i is governed by the balance of linear momentum in integral form

$$\rho_i \mathbf{a}_i = \int_{\mathcal{H}} \left(\mathbf{T}_i\langle\boldsymbol{\xi}\rangle - \mathbf{T}_j\langle-\boldsymbol{\xi}\rangle\right) d\boldsymbol{\xi} + \mathbf{g}, \tag{2}$$

where ρ_i is the density of particle i, $\mathbf{a}_i$ is its acceleration, $\mathbf{g}$ is the external body force, and the state function $\mathbf{T}$ models internal forces.

4 Viscoelastic Constitutive Model

We extend the classical elastic SBPD framework to incorporate viscoelastic behavior using a Prony-series-based energy model. Our approach captures time-dependent effects such as creep, relaxation, and hysteresis through control parameters. We implement our approach in a discrete numerical form that is compatible with particle-based simulations.

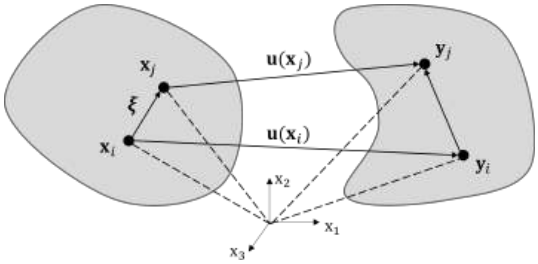


Fig. 1: Deformation state mapping.

The Prony model [18] is a widely used linear viscoelastic constitutive model which models the material's stress response σ as a sum of exponentially decaying functions via

$$\sigma(t) = E_\infty \cdot \varepsilon(t) + \sum_{k=1}^{N} E_k \cdot e^{-t/\theta_k} \cdot \varepsilon(t), \tag{3}$$

where ε is strain, N is the approximation order, E_∞ is the steady-state modulus, and E_k and θ_k are the relaxation modulus and relaxation time of the k-th mode, respectively.

To implement this model numerically, we discretize time and introduce variables q_k to capture the memory effect associated with each mode. These variables are updated over time as

$$q_k^{n+1} = \alpha_k \cdot q_k^n + (1 - \alpha_k) \cdot \varepsilon^{n+1}, \tag{4}$$

where $\alpha_k = e^{-\Delta t/\theta_k}$. Each q_k term gives the contribution of a specific relaxation mode and decays exponentially over time. This yields the stress update rule

$$\sigma^{n+1} = E_\infty \cdot \varepsilon^{n+1} + \sum_{k=1}^{N} E_k \cdot \left(\varepsilon^{n+1} - q_k^{n+1}\right). \tag{5}$$

Similar to the projected Peridynamics elastic model of by He et al. [6], we use a linear co-rotational elastic energy model to simulate the hyperelastic body and decompose it into a deviatoric part $\mathcal{W}^{\text{dev}}$ and an isotropic part $\mathcal{W}^{\text{iso}}$

$$\Psi = \int_{\mathcal{H}} w\langle\boldsymbol{\xi}\rangle \left(\mu\, \mathcal{W}^{\text{dev}}\langle\boldsymbol{\xi}\rangle + \frac{\lambda}{2}\, \mathcal{W}^{\text{iso}}\langle\boldsymbol{\xi}\rangle\right) d\boldsymbol{\xi}, \tag{6}$$

where μ and λ are the first and the second Lamé parameters, respectively. Assuming that all particles in $\mathcal{H}$ share the same deformation gradient $\mathbf{F}$, the ideal deformation tensor state can be expressed as $\hat{\mathbf{Y}} = \mathbf{F}\boldsymbol{\xi}$. $\mathcal{W}^{\text{dev}}$ is the energy of shear deformation and $\mathcal{W}^{\text{iso}}$ is the energy of volume deformation, which are defined as

$$\begin{aligned} \mathcal{W}^{\text{dev}} &= (|\hat{\mathbf{Y}}|/|\mathbf{X}| - 1)^2, \\ \mathcal{W}^{\text{iso}} &= (|\mathbf{Y}|/|\mathbf{X}| - 1)^2. \end{aligned} \tag{7}$$

When the horizon $\mathcal{H}$ is small and the deformation field is smooth, $\hat{\mathbf{Y}} \approx \mathbf{Y}$.

For each relaxation mode, we can now express the time evolution of the energy via our internal history variables as

$$\begin{aligned} \Psi_{i,j}^{\text{dev}} &= \mu_\infty \mathcal{W}_{i,j}^{\text{dev}} + \sum_{k=1}^{N} \mu_k \left(\mathcal{W}_{i,j}^{\text{dev}} - q_{i,j}^{\text{dev},k} \right), \\ \Psi_{i,j}^{\text{iso}} &= \frac{\lambda_\infty}{2} \mathcal{W}_{i,j}^{\text{iso}} + \sum_{k=1}^{N} \frac{\lambda_k}{2} \left(\mathcal{W}_{i,j}^{\text{iso}} - q_{i,j}^{\text{iso},k} \right), \end{aligned} \tag{8}$$

where, following (4), we have that

$$\begin{aligned} q_{i,j}^{\text{dev},k,n+1} &= \alpha_k q_{i,j}^{\text{dev},k,n} + (1-\alpha_k)\, \mathcal{W}_{i,j}^{\text{dev},n}, \\ q_{i,j}^{\text{iso},k,n+1} &= \alpha_k q_{i,j}^{\text{iso},k,n} + (1-\alpha_k)\, \mathcal{W}_{i,j}^{\text{iso},n}. \end{aligned} \tag{9}$$

The deviatoric force density is expressed as

$$\begin{aligned} \mathbf{T}_{ij}^{\text{dev}} &= \frac{2w(\boldsymbol{\xi})\gamma^{\text{dev}}}{|\mathbf{X}|^2}(|\mathbf{Y}| - |\mathbf{X}|)\text{dir}(\hat{\mathbf{Y}}), \text{ with} \\ \gamma^{\text{dev}} &= \mu_\infty + \sum_k \mu_k \left(1 - \frac{q_{i,j}^{\text{dev},k}}{\mathcal{W}_{i,j}^{\text{dev}}} \right). \end{aligned} \tag{10}$$

Similarly, the isochoric force density is given by

$$\begin{aligned} \mathbf{T}_{ij}^{\text{iso}} &= \frac{w(\boldsymbol{\xi})\gamma^{\text{iso}}}{|\mathbf{X}|^2}(|\mathbf{Y}| - |\mathbf{X}|)\text{dir}(\mathbf{Y}), \text{ with} \\ \gamma^{\text{iso}} &= \lambda_\infty + \sum_k \lambda_k \left(1 - \frac{q_{i,j}^{\text{iso},k}}{\mathcal{W}_{i,j}^{\text{iso}}} \right). \end{aligned} \tag{11}$$

In the above, γ is the effective modulus, *i.e.*, the effective stiffness of the deviatoric and isotropic components of the material at the current moment t. Using (10) and (11), we get the total force density $\mathbf{T}_{ij} = \mathbf{T}_{ij}^{\text{dev}} + \mathbf{T}_{ij}^{\text{iso}}$. Finally, we derive the discrete form of the equation of motion

$$\rho_i \mathbf{a}_i = h^2 \sum_{j \in \mathcal{H}} \left(\mathbf{T}_{ij}(\boldsymbol{\xi}) - \mathbf{T}_{ji}(-\boldsymbol{\xi}) \right) V_j. \tag{12}$$

5 Granular Material Simulation

Granular materials such as sand and snow often exhibit discrete elastoplastic behavior in the framework of continuum mechanics. We propose a peridynamics-based simulation method for granular flows under different yield criteria. We adopt the unified yield criterion proposed by Tu et al. [24] and implement three projection strategies for plastic mapping. Additionally, we dynamically update the Lamé parameters based on particle density to correct particle positions and enhance simulation stability.

5.1 Modeling yield for different materials

When a particle's internal stress state reaches the yield condition, irreversible plastic deformation occurs. We define a yield surface by the condition $\mathsf{y}(\boldsymbol{\tau}) \leq 0$, where $\boldsymbol{\tau}$ is the Kirchhoff stress tensor. If $\mathsf{y}(\boldsymbol{\tau}) > 0$, stress must be projected back to the yield surface, and the excess stress is interpreted as plastic flow. To define y, we first decompose the stress tensor $\boldsymbol{\tau}$ into

$$\mathbf{s} = \operatorname{dev}(\boldsymbol{\tau}),\ p = -\frac{1}{d}\operatorname{tr}(\boldsymbol{\tau}),\ q = \sqrt{\frac{6-d}{2}}\,\|\mathbf{s}\|, \tag{13}$$

where $d \in \{2, 3\}$ is the spatial dimension, $\mathbf{s}$ is the deviatoric stress tensor, the hydrostatic pressure p gives the compression or expansion of the volume, and the equivalent shear stress q gives the intensity of $\mathbf{s}$.

Granular materials and fluids: We model these by the Drucker–Prager yield criterion

$$\mathsf{y}_{\mathrm{vmdp}} = C_f \operatorname{tr}(\boldsymbol{\tau}) + \|\mathbf{s}\| - C_c = 0, \tag{14}$$

where C_f is the friction coefficient related to the friction angle, and C_c controls the intercept of the yield surface. When $C_f = 0$, the model degenerates into the Von Mises criterion, indicating purely shear-dominated yielding.

Clay and soil materials: We model these (under compressive loading) by the Cam-Clay yield criterion given by

$$\mathsf{y}_{\mathrm{vmcc}}(p, q) = C_f^2 p^2 + q^2 - C_c^2 = 0, \tag{15}$$

where C_f and C_c have similar meanings as in the Drucker–Prager model. C_c is the radius of the yield surface and is used to control hardening/softening behavior.

5.2 Plasticity mapping strategy

We simulate plastic deformation of granular materials such as sand or snow by implementing a plasticity mapping strategy within the SBPD framework. We use a classical 'return mapping' algorithm where plasticity is evolved by an elastic predictor step followed by a plastic corrector step: In the prediction step, plastic flow is temporarily ignored and stress and internal variables are updated elastically, yielding a trial deformation gradient $\mathbf{F}^{tr}$. If the yield surface is exceeded, we enter the plastic correction step and project stress back to the yield surface.

To incorporate plastic flow, we compute the elastic left Cauchy-Green deformation tensor as

$$\mathbf{b}^{tr} = \mathbf{F}_e^{tr}\mathbf{F}_e^{tr^T}. \tag{16}$$

Assuming a purely elastic response, the Kirchhoff stress tensor can be defined using a Neo-Hookean model as

$$\begin{aligned} \mathbf{s}^{tr} &= \mu J^{-2/d}\left(\mathbf{b}^{tr} - \frac{1}{d}\operatorname{tr}(\mathbf{b}^{tr})\mathbf{I}\right), \\ \boldsymbol{\tau}^{tr} &= \mathbf{s}^{tr} + \frac{\lambda}{2}(J^2 - 1)\mathbf{I}, \end{aligned} \tag{17}$$

where $J = \det(\mathbf{F}^{tr})$ is the volumetric change. The tensor $\mathbf{s}^{tr}$ captures shear response, while $\boldsymbol{\tau}^{tr}$ includes both volumetric and deviatoric effects.

We classify the return mapping into three cases (denoted A–C below) depending on the relation between the stress state and the yield limit $\tau_{\max} = C_c/C_f$.

Case A: If $\mathsf{y}(\boldsymbol{\tau}^{tr}) \leq 0$, stress lies inside the yield surface, thus $\mathbf{F}_p^{n+1} = \mathbf{F}^{tr}$.

Case B: If $|\operatorname{tr}(\boldsymbol{\tau}^{tr})| > \tau_{\max}$, the particle reaches the yield surface, setting $\boldsymbol{\tau}^{n+1} = \boldsymbol{\tau}_{\text{tip}}$. We update the principal stretch isotropically as

$$\begin{aligned} J^{n+1} &= \sqrt{\frac{2}{d\lambda}|\operatorname{tr}(\boldsymbol{\tau}_{\text{tip}})| + 1}, \\ \boldsymbol{\Sigma}^{n+1} &= (J^{n+1})^{1/d} \cdot \mathbf{I}, \end{aligned} \tag{18}$$

and compute the plastic deformation gradient using singular value decomposition

$$\mathbf{F}_p^{n+1} = \mathbf{U}\boldsymbol{\Sigma}^{n+1}\mathbf{V}^T. \tag{19}$$

This process represents the direct projection of stress to the yield apex under isochoric stretching, avoiding further decomposition in shear direction.

Case C: If $\mathsf{y}(\boldsymbol{\tau}^{tr}) > 0$ but the tip condition is not met, we perform a projection of the deviatoric stress norm. For the Drucker–Prager yield criterion, this becomes

$$\|\mathbf{s}^{n+1}\| = \|\mathbf{s}^{tr}\| - \mathsf{y}_{\text{vmdp}}(\boldsymbol{\tau}^{tr}). \tag{20}$$

For the Cam–Clay case, this becomes

$$\|\mathbf{s}^{n+1}\| = \sqrt{\|\mathbf{s}^{tr}\|^2 - \frac{2\mathsf{y}_{\text{vmcc}}(\boldsymbol{\tau}^{tr})}{6-d}}. \tag{21}$$

The deviatoric direction is preserved, and the updated stress is used to reconstruct the Cauchy-Green tensor:

$$\begin{aligned} \mathbf{s}^{n+1} &= \|\mathbf{s}^{n+1}\| \cdot \frac{\mathbf{s}^{tr}}{\|\mathbf{s}^{tr}\|}, \\ \mathbf{b}^{n+1} &= \frac{\mathbf{s}^{n+1}}{\mu J^{-2/d}} + \frac{1}{d}\operatorname{tr}(\mathbf{b})\mathbf{I}. \end{aligned} \tag{22}$$

The corrected plastic deformation gradient becomes

$$\mathbf{F}_p^{n+1} = \mathbf{U}\operatorname{diag}(\sqrt{\mathbf{b}^{n+1}})\mathbf{V}^T. \tag{23}$$

5.3 Dynamic adjustment of stiffness

In granular flow simulation, we no longer use fixed Lamé parameters, but instead update these adaptively based on local material compaction. Drawing from snow material handling methods in MPM [19], we estimate elastic response changes based on the particle's current compression density. We compute the local density as

$$\rho_i = \sum_j m_j W(\mathbf{x}_i - \mathbf{x}_j, h), \tag{24}$$

where W is a kernel function with support radius h. The local density reflects the current compression level of the material, and the rest density reads $\rho_{0,i}^t = \rho_i^t \left|\det(\mathbf{F}_{e,i}^t)\right|$.

Using the ratio of this rest density to the initial density, we dynamically adjust the current Lamé parameters as

$$\begin{aligned} \lambda_i^t &= \frac{E\nu}{(1+\nu)(1-2\nu)} \exp\left(\xi \cdot \frac{\rho_{0,i}^t - \rho_0}{\rho_{0,i}^t}\right), \\ \mu_i^t &= \frac{E}{2(1+\nu)} \exp\left(\xi \cdot \frac{\rho_{0,i}^t - \rho_0}{\rho_{0,i}^t}\right). \end{aligned} \tag{25}$$

This can be seen as a compression rate driven exponential hardening rule, which effectively enhances the response stiffness of materials such as snow in compacted states.

6 Boundary Handling

In the overall coupling of viscoelastic materials, granular flow materials, and rigid body boundaries, *boundary collision* mechanisms strongly influence simulation stability and realism. We introduce a boundary handling method using Sparse Signed Distance Fields (SDF) which improves stability and physical fidelity.

We directly sample and store SDF information on each rigid boundary particle, where each particle maintains a signed distance value ϕ and its gradient $\nabla\phi$, representing the shortest distance to the boundary and its direction, respectively. This design allows particle-to-particle collision detection and avoids repeated grid-based sampling. Collisions are triggered when the distance between particles is below a threshold $\|\mathbf{x}_i - \mathbf{x}_j\| < r$, or when $|\phi| < r$ for boundary contact.

Upon collision, particles are displaced along the contact normal direction with penetration depth $d = \min(|\phi_i|, |\phi_j|)$ and mass-based weighting. The contact normal is approximated by the gradient of the closer particle's SDF. For example, the position correction for particle i is given by:

$$\begin{aligned} \Delta\mathbf{x}_i &= -\frac{w_i}{w_i + w_j}(d \cdot \mathbf{n}_{ij}), \\ \Delta\mathbf{x}_j &= \quad \frac{w_j}{w_i + w_j}(d \cdot \mathbf{n}_{ij}), \end{aligned} \tag{26}$$

where $w_i = 1/m_i$.

To resolve sliding or sticking effects at boundaries, we introduce both dynamic and static friction models, as follows.

Dynamic friction: During particle-boundary contact, we compute the change in velocity due to collision $\Delta\mathbf{v}_i = \mathbf{v}_i^{n+1} - \mathbf{v}_i^*$, where $\mathbf{v}_i^{n+1}$ is the post-collision velocity and $\mathbf{v}_i^*$ is the elastic response velocity. We compute the tangential velocity as

$$\mathbf{v}_{it} = \mathbf{v}_i^{n+1} - \mathbf{n}v_{in}, \quad v_{in} = \mathbf{n} \cdot \mathbf{v}_i^{n+1}. \tag{27}$$

With $\mathbf{j} = m_i \Delta \mathbf{v}_i$ the impulse, the friction constraint reads $\|\mathbf{f}_t\| \leq c_b \|\mathbf{j}\|$. When the friction force can completely eliminate the tangential velocity, the velocity correction is simply $\mathbf{v}_i^{n+1} = \mathbf{n} v_{in}$. Otherwise we set

$$\mathbf{v}_i^{n+1} = \mathbf{v}_i^* - \frac{c_b}{m_i} \|\mathbf{j}\| \frac{\mathbf{v}_{it}}{\|\mathbf{v}_{it}\|}, \tag{28}$$

where c_b is the dynamic friction coefficient.

Static friction: To prevent persistent sliding near boundaries and simulate stacking behavior, we find stationary particles using a geometric criterion: If the motion of particle i satisfies

$$(\mathbf{y}_i^{t+1} - \mathbf{y}_i^*) \cdot (\mathbf{y}_i^t - \mathbf{y}_i^*) \geq \eta \|\mathbf{y}_i^* - \mathbf{y}_i^t\|^2, \tag{29}$$

we freeze its position, *i.e.*, set $\mathbf{y}_i^{t+1} = \mathbf{y}_i^t$. η is the static friction coefficient, set to $\eta = 0.8$ in our simulations.

7 Results and Discussion

We implemented our framework on an NVIDIA GeForce RTX 4090 GPU using the Taichi programming language for efficient parallel simulation. The overall simulation procedure is outlined in Algorithm 1, where we typically set the maximum number of iterations $\text{iter}_{\max}$ to 5, and terminate early if the maximum iteration displacement falls below a predefined threshold $\epsilon = 10^{-4}$. All visual results were rendered offline via Houdini. Detailed simulation performance information is given in Table 1.

Table 1: Simulation information for selected examples. P is the number of particles.

Exp.	P	Δt	FPS	E_0	ν
Fig. 2	80k	5 ms	68.67	1×10^8	0.45
Fig. 3	195k	2 ms	27.20	1×10^7	0.45
Fig. 4	348k	2 ms	13.19	3×10^5	0.20
Fig. 5	95k	5 ms	17.06	2×10^5	0.20
Fig. 6	167k	2 ms	7.56	2×10^5	0.20
Fig. 7	416k	1 ms	8.80	1×10^7 (elast.) 2×10^5 (sand)	0.25 0.20

Viscoelastic stretch: We validate our algorithm using a $N = 3$ (rd) order Prony model. The total Young's modulus E_0 gives the initial stiffness of the material, while the long-term modulus E_∞ characterizes its stiffness at infinite time. Each E_k denotes the relaxation modulus of the k-th component, with θ_k being the corresponding relaxation time. The material behavior is defined using

Algorithm 1 Elastomer-Sand Coupling Simulation Based on SBPD

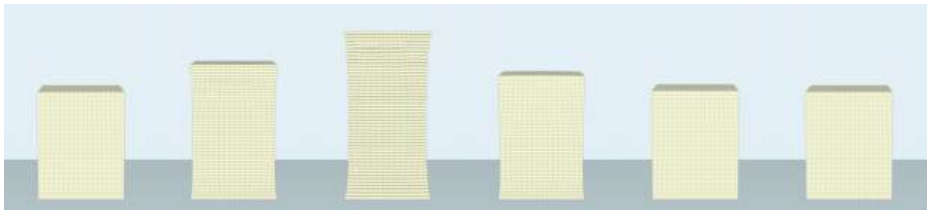
Input: $\mathbf{y}^t, \mathbf{v}^t$, phase, Δt, $\text{iter}_{\max}$, E_0, ν, E_∞, E_k, θ_k, C_f, C_c, η, ϵ
Particle advection: $\mathbf{y}^{t+1} \leftarrow \mathbf{y}^t + \mathbf{v}^t \Delta t$
while iteration $<$ $\text{iter}_{\max}$ and $\max(\|\Delta \mathbf{x}_i\|) > \epsilon$ **do**
    // **Elastic Phase:**
    Compute deformation gradient $\mathbf{F}$
    Compute force density $\mathbf{T}$ (Eq. (10), (11))
    Compute displacement $\Delta \mathbf{x}$ (Eq. (12))
    Update position: $\mathbf{y}^{t+1} \leftarrow \mathbf{y}^{t+1} + \Delta \mathbf{x}$
    // **Sand Phase:**
    Compute deformation gradient $\mathbf{F}$
    Project $\mathbf{F}$ onto yield surface (Cases A/B/C)
    Compute force density $\mathbf{T}$
    Compute displacement $\Delta \mathbf{x}$ (Eq. (12))
    Update position: $\mathbf{y}^{t+1} \leftarrow \mathbf{y}^{t+1} + \Delta \mathbf{x}$
end while
// **Constraints and Collisions:**
while iteration $<$ $\text{iter}_{\max}$ **do**
    Apply self and inter-phase collision response (Eq. (26))
    Apply boundary advection
end while
// **Post-processing:**
Update velocity: $\mathbf{v}^{t+1} \leftarrow \left(\mathbf{y}^{t+1} - \mathbf{y}^t\right)/\Delta t$
Apply static friction constraint
Apply dynamic friction constraint
Update neighbor list j
Update Lamé parameters (Eq. (25))

the empirical relation: $E_0 = E_\infty + \sum_{k=1}^{N} E_k$. The configuration of relaxation times at each order can be determined according to the empirical rules of exponential decay.

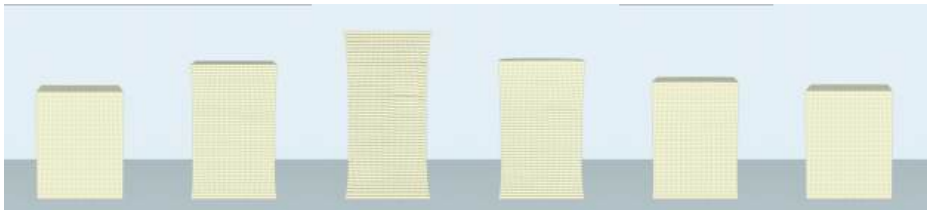
Figure 2 shows a stretching–unloading experiment that compares the relaxation behavior of hyperelastic, viscoelastic, and elastoplastic materials after external force removal. The hyperelastic model was configured with $E_0 = E_\infty$ and recovered quickly upon unloading, with almost no energy dissipation. For the viscoelastic model, we set $E_\infty = 0.4E_0$, $E_k = [0.3, 0.2, 0.1] \cdot E_0$, and $\theta_k = [0.1, 1.0, 5.0]$. The recovery behavior showed exponential time-decay characteristics (see the supplemental video). The elastoplastic model with Von Mises yield criterion showed significant energy dissipation and permanent deformation.

We further illustrate the flexibility of our viscoelastic model by an "armadillo stretch-rest-unload" experiment with $E_0 = 1 \times 10^7$ and $\nu = 0.45$. We compared three different viscoelastic material parameters:

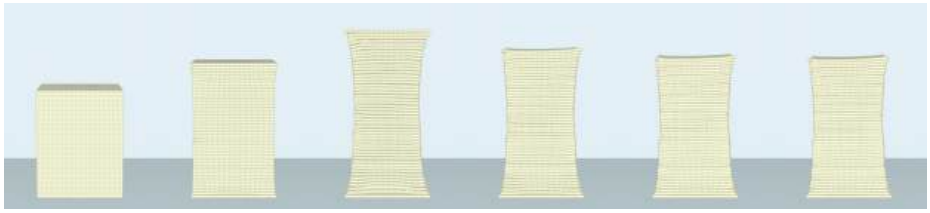
- Purely elastic: $E_\infty = E_0$;
- High viscosity: $E_\infty = 0.3E_0$, $E_k = [0.3, 0.2, 0.2] \cdot E_0$, $\theta_k = [0.5, 2.0, 5.0]$;
- Low viscosity: $E_\infty = 0.5E_0$, $E_k = [0.25, 0.15, 0.1] \cdot E_0$, $\theta_k = [0.5, 2.0, 5.0]$.



(a) Stretch-release process of pure elastic body [6]



(b) Stretch-release process of viscoelastic body



(c) Stretch-release process of elastoplastic body

Fig. 2: Comparison of recovery behavior of different materials during the stretch-release process.

Figure 3 shows the evolution of energy (in red) throughout the simulation and presents a quantitative analysis. During the stretching phase, the total energy of all three materials increases non-linearly due to the work done by external forces, consistent with the non-linear characteristics of stress-strain relationships. For the purely elastic material, the external work is entirely converted into elastic potential energy, whereas for viscoelastic materials, a portion of the energy is dissipated through viscous effects. In the constant-stretching phase, the energy of the elastic material remains unchanged, while the viscoelastic materials exhibit stress relaxation, demonstrating the physical plausibility of our model. In the relaxation phase, the purely elastic material released energy most rapidly and almost completely returned to its original state. Highly viscous materials release energy more slowly, showing significant hysteresis effects as part of the energy is converted to heat through viscous mechanisms. The energy release rate of the low-viscosity elastic material lies between the two. These results demonstrate the effectiveness of our viscoelastic model and its strong tunability in capturing diverse material responses.

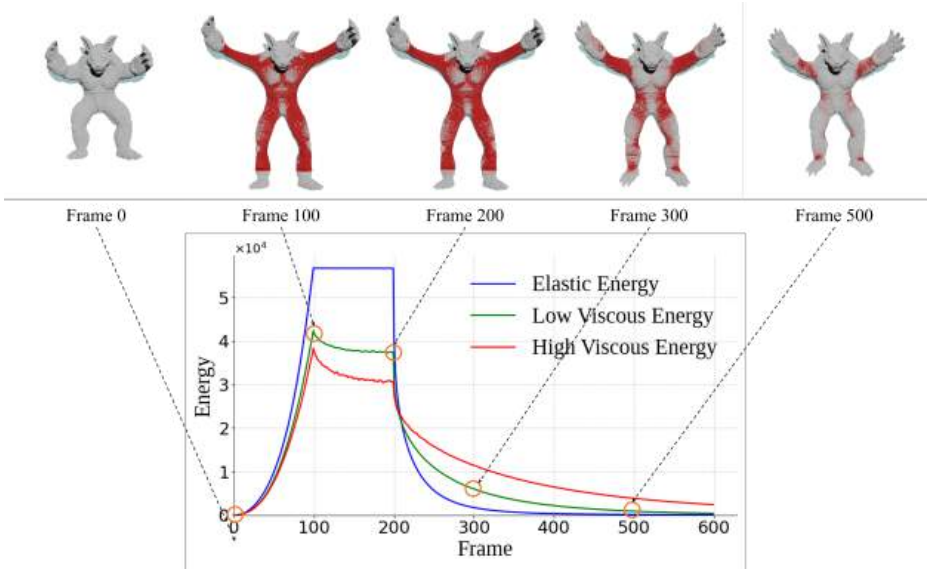


Fig. 3: Energy evolution of the armadillo stretching experiments.

Sand simulation: We designed a series of granular flow experiments and compared them with MPM simulations based on the Drucker–Prager yield criterion. These comparisons validate the effectiveness of different yield mapping schemes under the peridynamic framework in reproducing physically plausible granular flow and pile-up behaviors.

Figure 4 shows a sand pile experiment with $E_0 = 3 \times 10^5$ and $\nu = 0.2$. Under high friction coefficients, our method successfully produced stable, high-friction sandpiles in which the upper particles resisted sliding. Compared to the MPM approach under the same friction angle and coefficient, our method achieved more pronounced pile-up effects by introducing stronger cohesive forces. Additionally, the peridynamics framework, extended from elastic energy, allows for larger time steps, improving overall simulation efficiency.

To further study the influence of cohesion, we conducted a slope-divided sand pile experiment (Fig. 5). We used the Drucker-Prager yield criterion with $E_0 = 2 \times 10^5$ and $\nu = 0.2$, and used materials with different cohesion coefficients. Under higher cohesion, some sand particles could adhere to the inclined surface forming local accumulations. For lower cohesion, only a thin layer of particles remained, with the rest quickly sliding down. The accumulation patterns on the ground also showed significant differences: high-cohesion materials formed more compact sand pile structures; low-cohesion materials appeared more dispersed.

To evaluate the influence of friction coefficients on granular flow behavior and accumulation patterns, we conducted an hourglass experiment under constant cohesion (Fig. 6). The results show that higher friction coefficients yield in poorer flow of particles near boundaries, while internal particles still show a certain flow.

(a) Sand piling based on MPM [7]

(b) Sand piling under high friction coefficient

(c) Sand piling under high cohesion coefficient

Fig. 4: Sand piling simulation experiments.

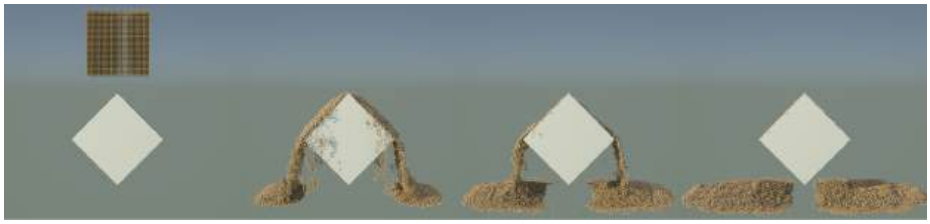
In contrast, materials with low friction exhibited more uniform flow between interior and exterior regions. After exiting the funnel, high-friction materials formed taller and steeper piles with an angle of repose measuring 30.61°, while low-friction materials produced flatter deposits with an angle of repose of 20.80°.

Coupling simulation: To validate the multi-material coupling capability of our framework, we performed an experiment involving viscoelastic bunnies interacting with bunny-shaped sand (Fig. 7). The viscoelastic material has $E_0 = 1 \times 10^7$, $\nu = 0.25$. Sand particles have $E_0 = 2 \times 10^5$, $\nu = 0.2$. During free fall, the viscoelastic bunnies undergo deformation upon impact, while sand flows into the gaps between them and forms a stable pile. The experiment shows realistic two-way coupling, where both material types influence each other's behavior under collision and accumulation.

8 Conclusions and Future Work

We proposed a unified visco-elasto-plastic simulation framework based on SBPD to address the limitations of CCM in modeling discontinuities. Our framework demonstrates flexibility and effectiveness in simulating both viscoelastic solids and granular materials.

In terms of viscoelastic simulation, we derived time-dependent force density formulations based on the Prony model, accurately capturing complex response



(a) Cohesion coefficient $C_c = 1$

(b) Cohesion coefficient $C_c = 100$

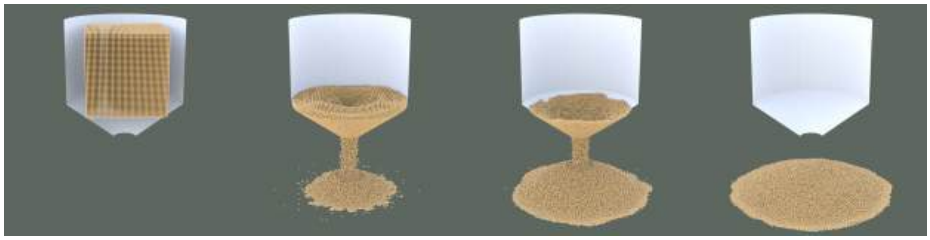
Fig. 5: Sand-slope experiments.

characteristics such as stress relaxation, creep, and hysteresis. For granular flow simulation, we integrated various yield criteria and mapping strategies, combined with density-based dynamic stiffness adjustment mechanisms, achieving natural flow, accumulation, and separation behaviors of particles. The framework further supports interactions among viscoelastic solids, granular media, and rigid bodies via a multi-material coupling mechanism, enhancing its robustness and applicability.

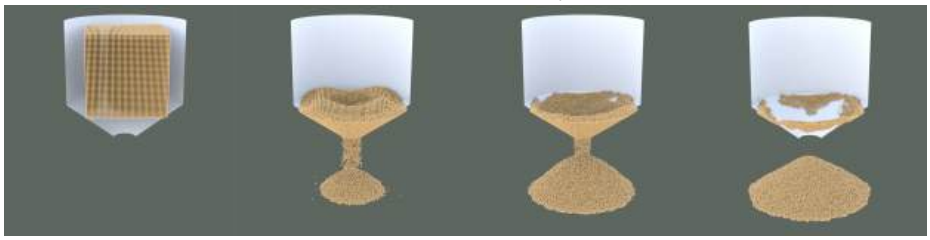
However, the computational efficiency of the current method for large-scale granular flow simulations still remains a challenge. In future work, we will focus on developing implicit iterative acceleration strategies to enhance the stability and efficiency of large-scale computations. Furthermore, we plan to leverage the advantages of Peridynamics in handling fracture and crack propagation by incorporating fracture mechanics mechanisms into the viscoelastic model, enabling the simulation of richer material discontinuity behaviors. Building on the extensibility of our framework, we will also integrate viscoelastic fluids into the unified framework, extending its application capabilities in biological fluid and soft matter simulations.

Acknowledgments. This study was funded by National Major Science and Technology Project of China (2024ZD0608100), National Natural Science Foundation of China (Nos.62376025, 62332017), Guangdong Basic and Applied Basic Research Foundation (No.2023A1515030177), China Scholarship Council(No.202306460015).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.



(a) Friction coefficient $C_f = 0.1$



(b) Friction coefficient $C_f = 2$

Fig. 6: Hourglass experiments.

Fig. 7: Viscoelastic bunny and sand coupling experiment.

References

1. Bußler, M., Diehl, P., Pflüger, D., Frey, S., Sadlo, F., Ertl, T., Schweitzer, M.A.: Visualization of fracture progression in peridynamics. Computers & Graphics (2017). https://doi.org/https://doi.org/10.1016/j.cag.2017.05.003, http://www.sciencedirect.com/science/article/pii/S0097849317300456
2. Daviet, G., Bertails-Descoubes, F.: A semi-implicit material point method for the continuum simulation of granular materials. ACM TOG **35**(4), 1–13 (2016)
3. Fang, Y., Li, M., Gao, M., Jiang, C.: Silly rubber: an implicit material point method for simulating non-equilibrated viscoelastic and elastoplastic solids. ACM TOG **38**(4), 1–13 (2019)

4. Gao, M., Pradhana, A., Han, X., Guo, Q., Kot, G., Sifakis, E., Jiang, C.: Animating fluid sediment mixture in particle-laden flows. ACM TOG **37**(4), 1–11 (2018)
5. Gissler, C., Henne, A., Band, S., Peer, A., Teschner, M.: An implicit compressible sph solver for snow simulation. ACM TOG **39**(4), 36–1 (2020)
6. He, X., Wang, H., Wu, E.: Projective peridynamics for modeling versatile elasto-plastic materials. IEEE TVCG **24**(9), 2589–2599 (2017)
7. Klár, G., Gast, T., Pradhana, A., Fu, C., Schroeder, C., Jiang, C., Teran, J.: Drucker-Prager elastoplasticity for sand animation. ACM TOG **35**(4), 1–12 (2016)
8. Lai, X., Ren, B., Fan, H., Li, S., Wu, C., Regueiro, R., Liu, L.: Peridynamics simulations of geomaterial fragmentation by impulse loads. Intl J Num Analy Meth Geomech **39**(12), 1304–1330 (2015)
9. Lenaerts, T., Dutré, P.: Mixing fluids and granular materials. In: Comp Graph Forum. vol. 28, pp. 213–218 (2009)
10. Madenci, E., Oterkus, S.: Ordinary state-based peridynamics for thermoviscoelastic deformation. Eng Fract Mech **175**, 31–45 (2017)
11. Müller, M., Charypar, D., Gross, M.: Particle-based fluid simulation for interactive applications. In: Proc. SCA. pp. 154–159 (2003)
12. Narain, R., Golas, A., Lin, M.: Free-flowing granular materials with two-way solid coupling. In: Proc. SIGGRAPH Asia, pp. 1–10 (2010)
13. Ozdemir, M., Oterkus, S., Oterkus, E., Amin, I., El-Aassar, A., Shawky, H.: Fracture simulation of viscoelastic membranes by ordinary state-based peridynamics. Proced Struct Integr **41**, 333–342 (2022)
14. Peer, A., Gissler, C., Band, S., Teschner, M.: An implicit sph formulation for incompressible linearly elastic solids. Comp Graph Forum **37**(6), 135–148 (2018)
15. Romeo, M., Monteagudo, C., Sánchez-Quirós, D.: Muscle and fascia simulation with extended position based dynamics. Comp Graph Forum **39**(1), 134–146 (2020)
16. Silling, S.: Reformulation of elasticity theory for discontinuities and long-range forces. J. Mech Phys Solids **48**(1), 175–209 (2000)
17. Silling, S., Epton, M., Weckner, O., Xu, J., Askari, E.: Peridynamic states and constitutive modeling. J. Elast **88**, 151–184 (2007)
18. Simo, J.C., Hughes, T.J.: Computational inelasticity, vol. 7. Springer Science & Business Media (2006)
19. Stomakhin, A., Schroeder, C., Chai, L., Teran, J., Selle, A.: A material point method for snow simulation. ACM TOG **32**(4), 1–10 (2013)
20. Takahashi, T., Dobashi, Y., Fujishiro, I., Nishita, T., Lin, M.: Implicit formulation for sph-based viscous fluids. Comp Graph Forum **34**(2), 493–502 (2015)
21. Tampubolon, A., Gast, T., Klár, G., Fu, C., Teran, J., Jiang, C., Museth, J.: Multi-species simulation of porous sand and water mixtures. ACM TOG **36**(4), 1–11 (2017)
22. Terzopoulos, D., Fleischer, K.: Modeling inelastic deformation: viscolelasticity, plasticity, fracture. In: Proc. CGI. pp. 269–278 (1988)
23. Terzopoulos, D., Platt, J., Barr, A., Fleischer, K.: Elastically deformable models. In: Proc. CGI. pp. 205–214 (1987)
24. Tu, Z., Li, C., Zhao, Z., Liu, L., Wang, C., Wang, C., Qin, H.: A unified MPM framework supporting phase-field models and elastic-viscoplastic phase transition. ACM TOG **43**(2), 1–19 (2024)
25. Viriyayuthakorn, M., Caswell, N.: Finite element simulation of viscoelastic flow. J. Non-Newtonian Fluid Mech **6**(3-4), 245–267 (1980)
26. Yue, Y., Smith, B., Batty, C., Zheng, C., Grinspun, E.: Continuum foam: A material point method for shear-dependent flows. ACM TOG **34**(5), 1–20 (2015)

27. Zhu, K., He, X., Li, S., Wang, H., Wang, G.: Shallow sand equations: real-time height field simulation of dry granular flows. IEEE Transactions on Visualization and Computer Graphics **27**(3), 2073–2084 (2019)
28. Zhu, Y., Bridson, R.: Animating sand as a fluid. ACM TOG **24**(3), 965–972 (2005)

Immersion Discrepancies in Educational Serious Games Among Children's Age Groups

Hui Liang[1*], Yukun Li[1], JiaLin Fu[1]

[1] Zhengzhou University of Light Industry, 136 Science Avenue, Zhengzhou 450001, Henan, China
hliang@zzuli.edu.cn

Abstract. With the development of virtual reality technology, serious games have become a new type of teaching tool, and exploring the differences in their sense of immersion is of great significance in enhancing user experience and promoting personalized education. In this study, we designed three educational-themed serious games and compared the power spectral densities (PSD) of immersion-related brain waves of children of different ages by using a difference analysis algorithm based on the game test model. The results showed that the PSDs of theta, alpha, and beta waves differed significantly in different age groups; in the tutor-guided experiment, only theta wave differed significantly. The younger group had higher levels of θ-wave and α-wave activity, and were more relaxed and creative during the game; the older children had higher levels of β-wave activity, and had better attention and cognitive level during the game. This study reveals the influence of age on children's cognitive and emotional participation in educational games from a neurophysiological point of view, and provides a neuroscientific basis for the development of personalized educational tools.

Keywords: VR education, serious game, EEG, immersion.

1 Introduction

Serious games have gained widespread adoption in educational settings for school-age children, as they bolster the quality of the learning experience and academic performance. They offer a platform for knowledge exchange, collaborative learning, and social interaction [1-3]. When integrated with traditional educational resources, serious games offer unique visualization and interaction prospects [4], maintaining a high level of motivation, which, in turn, augments the overall learning experience [5-6].

Successful serious games share a common trait: their capacity to engross players in an immersive experience. This phenomenon is commonly referred to as “immersion.” To delve into the determinants of immersion, Brown et al. [7] conducted a qualitative study, confirming immersion as a descriptor of a player's engagement level in a game. The greater the engagement, the deeper the immersion, and correspondingly, the player's emotional responses are profoundly influenced by the game's immersive

qualities, immersive environments lead to more positive emotions. Numerous studies emphasize the potential of serious games to augment learning by boosting motivation, thus leading to enhanced learning outcomes [8]. Barclay et. al. [9] established a correlation between immersion and improved learning outcomes. Beyond its educational benefits, serious games serve as potent motivators in student education [10] while fostering the development of cognitive skills such as problem-solving, creativity, and critical thinking [11].These advantages extend even to students prone to inattention [12].Moreover, serious games facilitate the acquisition of skills including discovery-based learning [13],motor skills, spatial coordination [14],and expertise development [15].Serious game affected positively the children's basic learning mechanisms (BLMs), by reinforcing balance, visual-motor, memory, attention, and spatial awareness abilities while interacting with the serious game.[16].

Despite the availability of AI tools [17] and game design frameworks tailored to serious game design [18-19], these resources remain insufficient in offering comprehensive guidance for the incorporation of immersion elements into serious games. In addition, there exists a paucity of in-depth studies concerning the variability of physiological markers of immersion in serious games for school-aged children, who constitute the primary demographic of serious game users.

This study combines neuroscience and education to investigate school-aged children's immersion in serious games. Part 1 presents the research background and objectives. Part 2 reviews relevant literature, focusing on correlation studies. Part 3 details the experimental design, covering the differential analysis algorithm based on the proposed serious game test model, EEG data collection methods, and game design. Part 4 describes the experimental procedures and data processing. Part 5 discusses the results, and Part 6 concludes the study.

2 Related Work

Serious games are experiencing rapid growth and are extensively employed in children's education. Cheng et al. [20] discovered a significant correlation between learning outcomes in educational games and the subjective experience of immersion. Achieving a balance between the effectiveness of serious games and the enjoyment of the experience poses a challenging task in contemporary game design. Investigating the emotional dynamics of players in the game environment has been proposed as an effective solution.

Barclay and Bowers [21] observed that the benefits of immersion in serious educational games are no longer solely attributable to highly available systems or exceptionally receptive learners. While some studies have applied game design principles to the educational process in serious games, there remains still a significant absence of systematic and empirically tested design methodologies [22]. Additionally, research has explored the variability in experiential perception and acceptance across different age groups in the context of serious gaming experiences. For instance, some researchers [23] assessed the performance and subjective experiences of three age groups in serious gaming, revealing significant differences in re-gaming experiences

and processing speed among these age groups. Chiang et al. [24] created an EEG-based model to objectively gauge attention and learning capacities. Moreover, Wan et al. conducted an assessment of immersive learning in university students, providing evidence supporting the feasibility of predicting the level of learning immersion through physiological recordings.

The study focused on the child population's cognitive abilities, attention, and expression. Due to the unsuitability of children for questionnaire participation, subjective evaluations of the children were not collected. To ensure an accurate and objective assessment, brainwaves were utilized to physiologically measure the participants' brain activity. Previous research by scholars has assessed the brainwaves of individuals engaged in serious games. Alpha waves (18-12 Hz) are known to play a significant role in various sensory and cognitive processes and exhibit a negative correlation with attention [25]and cortical activation [26]. Beta band oscillations (15-30 Hz) have been proposed as indicators of cognitive processing, particularly in the upper part of the beta band. Theta EEG bands (4-7 Hz) have been linked to memory and cognitive abilities. In a study conducted by Škola et al. [27] on presence, engagement, and immersion in virtual reality, it was found that the total duration of a VR application was inversely related to technology adoption and negatively correlated with immersion. This suggests a negative association between the duration of VR application and the levels of presence, engagement, and immersion in virtual reality. These findings emphasize the potential of EEG as a viable and objective method for assessing immersion [28]. In terms of brainwave frequencies, alpha waves are consistently recorded and are sensitive to changes in task difficulty. They are the dominant waves in human EEG brain recordings in the range of 7.5-13.5 Hz. The evaluation of immersion comprises various aspects, including perception, control, attention, enjoyment, and self-awareness [29-32]. Indicators of immersion in different brain waves are shown in Fig. 1.

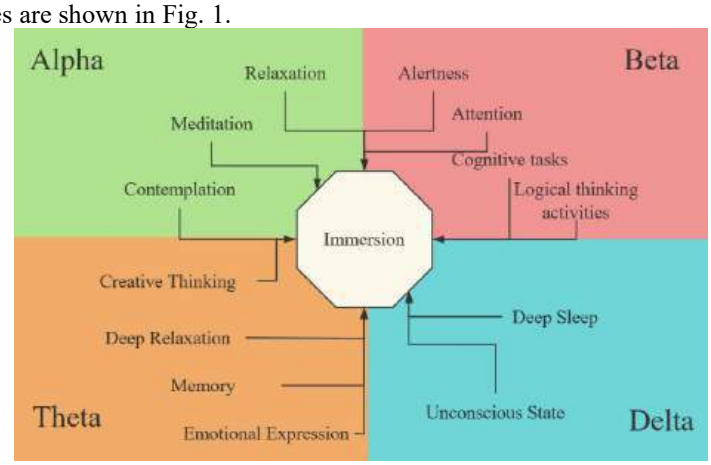


Fig. 1. Indicators of immersion in different brain waves

In summary, numerous studies have highlighted the benefits of serious games in enhancing children's learning abilities and immersive learning experiences. However, there remains a significant gap in research when it comes to understanding the differences in immersion at the physiological level in serious educational games. This

gap is primarily due to differences in cognitive development, emotional responses, and levels of concentration among children of various age groups. There is a clear need for more comprehensive analyses into EEG indicators of immersion. To address this need, this study has developed a disparity analysis algorithm based on a serious game testing model. The primary objective is to verify the differences in immersion levels experienced by children of different age groups while engaging with serious educational games. To achieve this objective, we have designed three serious educational games with educational themes, which will be utilized to assess the participating children.

3 Research Design

3.1 Design of Variance Analysis Algorithm Based on Serious Game Testing Models

To investigate differences in the level of immersion among children of varying ages when engaging with serious games, a Variance Analysis algorithm is proposed, utilizing the Serious Game Test model. In this analytical framework, age serves as the primary independent variable. Employing the Serious Game Test model enables us to identify differences not only between different age groups but also in each age group. Our study categorizes children into high and low age groups, with their data hierarchically nested. Failing to acknowledge this nested relationship and conducting a simple comparison might result in an oversight of the relationship between individual and group data, leading to imprecise difference estimations. To address these potential issues, our proposed model, grounded in serious game testing, adeptly tackles the matter. It not only scrutinizes differences between groups but also within them, while also controlling for potential interfering factors.

The Variability Analysis algorithm, rooted in the Serious Play Test model, evaluates the variability in mean EEG signals during serious play among children. The inferences drawn from this analysis are based on certain assumptions. Before performing the Variability Analysis, it is necessary to subject the mean power values of the EEG signals, obtained from processing, to normality and chi-square tests.H0 represents the null hypothesis, signifying the assumption that no effect or difference exists in the overall parameters or distribution. In the context of analysis of variance, H0 posits that no difference exists. On the contrary, H1 signifies the alternative hypothesis, representing the opposite of the null hypothesis, often suggesting the presence of a difference. Prior to the analysis of variance, normality and chi-square tests are conducted, establishing hypothesis tests using both H0 and H1.

The statistical D-value is employed for making inferences and evaluating the significance of differences. This statistic compares differences between groups with differences in groups. In the framework of analysis of variance, a comparison is necessary between between-group differences, which reflect differences in group means, and in-group differences, which indicate the degree of variability in observations of each group. D-values are computed by contrasting between-group variance with in-group variance to measure the magnitude of between-group

differences relative to in-group variance. Therefore, larger D-values indicate a significantly greater between-group variance compared to in-group variance, and vice versa. At the conclusion of the D-value analysis, it becomes essential to determine the criticality and make a statistical decision. Criticality pertains to the reference value employed in analysis of variance to determine the significance of the D-value. Statistical software, such as SPSS, is utilized to establish criticality and reach a statistical decision. To gain a comprehensive understanding of discrepancies in EEG metrics among children, taking into account EEG data's complexity and noise, repeated comparisons are conducted. This approach yields more exhaustive, profound, and stable findings. The following is the procedural outline:

(1) Statistical D-value

Means were calculated first: it represents the sample mean and the total mean for the ith overall level, with ni representing the number of sample observations for the ith overall level. Sum of Squared Errors: Calculate the sum of squared errors, which comprises the between-group sum of squares (Sa). Sa reflects the extent of difference between sample means of the overall levels and indicates the impact of differences in theoretical means of factor A. This is labeled as “the sum of squares of factor A” or “between-group difference.”

In-Group Sum of Squares: it represents the in-group sum of squares (Se). Se is the sum of squared errors between the sample data of each group and its group mean, illustrating the dispersion of each observation in each sample and denoting the effect of random errors. It is referred to as “sum of squares of errors” or “in-group variance”.

Total error sum of squares St: it represents the sum of the squares of errors across all observations and the overall mean, serving as an indicator of the dispersion among all observations. The between-group and in-group mean squares are obtained by dividing the sum of squared errors by their respective degrees of freedom. The Ma/Me ratio forms the basis of the D distribution.

(2) Critical value is determined and statistical decisions are made

After calculating the D statistic, one should locate the corresponding critical value, Alpha, in the D distribution table. This is achieved for a numerator with degrees of freedom of (k-1) and a denominator with degrees of freedom of n-k, according to the given significance level, Alpha.

When the value of D is greater than the critical D value, it is indicative of rejecting the null hypothesis in favor of the alternative hypothesis, supported by our data.

When the value of D is less than the critical D value, it is not advisable to conclude the acceptance of the null hypothesis. Instead, it is more appropriate to state that the null hypothesis was not rejected.

(3) Multiple comparisons

The difference between individual EEG indicator point estimates plays a role in reinforcing the conclusion mentioned above. If this difference is not sufficient to be practically significant, it further emphasizes that any existing differences between levels, if present, hold limited practical importance.

Should the difference in mean values between different levels reach a level of significance from an applied perspective, the original hypothesis H0 is accepted due

to the significant effect of random error. Conversely, if this value is considered excessively large from an applied standpoint, it suggests that the present test lacks precision. In such cases, it is advisable to consider increasing the test size and improving the test to minimize the effect of random errors.

(4) Algorithm design

The sum of squares between age groups, denoted as Sa, is the sum of the squares of errors between the group means and the overall mean. This reflects the extent of difference among the sample means at each level of aggregation and signifies the effect of differences in the theoretical mean at each level of Factor A. It is also referred to as the "sum of squares of Factor A" or the "between-group difference." The calculation process of Sa is shown in Formulate.1.

$$\mathrm{Sa} = \sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{ni}(\overline{\mathrm{x}_{\mathrm{i}}} - \overline{\overline{\mathrm{x}}})^2 \tag{1}$$

On the other hand, the in-group sum of squares, Se, comprises the sum of the squares of errors in the sample data of each group and its respective group mean. This reflects the dispersion of each observation in each sample and indicates the effect of random errors. It is denoted as the "sum of squares of errors" or "in-group variation." The calculation process of Se is shown in Formulate.2.

$$Se = \sum_{i=1}^{k} \sum_{j=1}^{ni} \left(x_{ij} - \overline{x_i}\right)^2 \tag{2}$$

The total error sum of squares, St = Se+ Sa, represents the sum of the squares of errors across all observations and the overall mean, reflecting the dispersion among all observations. The calculation process of St and D-value are shown in Formulate.3 and 4.

$$St = \sum_{i=1}^{k} \sum_{j=1}^{ni} (x_{ij} - \overline{\overline{x}})^2 \tag{3}$$

$$D = \frac{Ma}{Me} = \frac{Sa}{k-1} \Big/ \frac{Se}{n-k} = \frac{Sa \times Se}{(k-1)(n-k)} \sim F(k-1, n-k) \tag{4}$$

3.2 Design of Spectral Analysis

In this study, spectral features of electroencephalogram (EEG) signals are derived from four frequency bands: delta (0-3.5 Hz), theta (3.5-7.5 Hz), alpha (7.5-13.5 Hz), and beta (13.5-26 Hz). Given that EEG signals exhibit non-stationary characteristics over short periods, the Discrete Wavelet Transform (DWT) is adopted for feature extraction, as it outperforms the Fast Fourier Transform (FFT) in this context [32]. DWT utilizes scale and wavelet functions associated with low-pass and high-pass filters respectively. By passing the original signal X[n] through these filters and applying the Nyquist sampling rule to discard half of the samples, the signal is

decomposed into different frequency bands. This subband coding process can be iterated, reducing time resolution by half while doubling the frequency resolution at each level, enabling detailed analysis of the signal across various resolutions and frequency bands.

Raw EEG data captured by the EEG device often contains artifacts from muscle activity, eye movements, and heart rate variability. To address this, BESA software is employed to remove these artifacts, yielding task-relevant raw signals. Subsequently, the targeted EEG waves are accurately extracted for further analysis.

3.3 Serious Game Design

We designed three serious games for elementary school age children. Among them, “Jing Ke Stabbing Qin” and “Grass Boat Borrowing Arrows” are inspired by Chinese culture, and “The Crow and the Water Jar” is inspired by Aesop's fables. The game is based on the story of Zhuge Liang's borrowing of arrows, in which the player has to control a thatched boat to collect 300 arrows and avoid bombs. The game adopts a cartoon interface, and is set up with three types of weather to increase the difficulty: sunny, rainy, and foggy, with gestures required to control the game in rainy and foggy days. Jing Ke Qin” reproduces the plot of Jing Ke's assassination of Qin by manipulating the shadow characters and utilizing five control points and depth sensors to complete the game. In “The Crow and the Water Jar”, children have to use gestures to manipulate the puppet crow to pick up pebbles and put them into the water jar, and the game process provides feedback to assist in the measurement of EEG metrics. The games involved are shown in Fig. 2.

In each game, two modes of experience were set up, the lower age group was guided by an adult tutor, and the upper age group completed the game independently, in order to explore the differences in brain waves between the guided and unguided brain waves of different age groups, to understand the electroencephalographic mechanism of game immersion, and to compare the immersion indexes of the two groups, to explore the relationship between age and immersion.

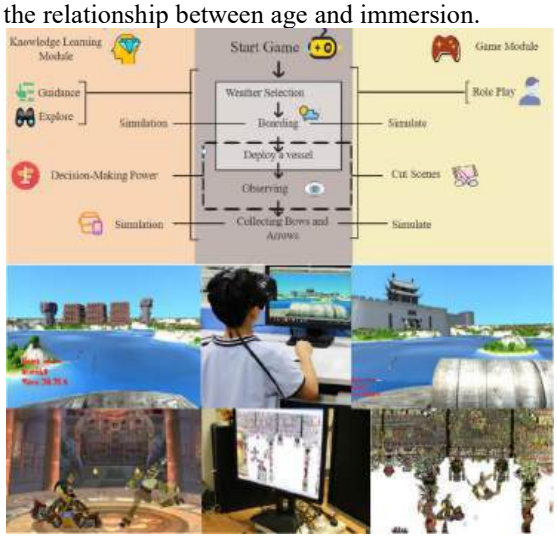


Fig. 2. Serious game design

4 Materials and Methods

4.1 Participants

The participants were 64 children of different school ages. According to the age division of primary school-aged children classified by the Ministry of Education of the People's Republic of China, our recruitment interval was set at 6-11 years old, and the participants were divided into a high age group (Group H) and a low age group (Group L) according to their age, with Group H's age range being 6-8 years old and Group L's age range being 9-11 years old. The age distribution of the participants was close to normal distribution to ensure the representativeness and reliability of the results.

4.2 Data collection and extraction

Prior to the experiment, children's parents will receive an information sheet and consent form containing detailed information about the purpose of the study, the game design, and the type of EEG data, and sign it after one week's consideration. The experiment follows the guidance of the UCLM Research Ethics Committee, which is in line with the ethical requirements of children's research.

The experiment was conducted in the Serious Play Experience Laboratory, where children were accompanied by their parents and entered for ten minutes of interaction to familiarize themselves with the environment, followed by a three-minute break. After the children were relaxed, an immersion-inducing experiment was conducted, experiencing two identical games with a three-minute break between games to calm down. EEG signals were collected in a soundproof, electrically shielded room using a 60/64 EEG device (emotiv epoc X). Both upper and lower age groups experienced the three games, with the lower age group being guided by a tutor for Crow and the Water Jar (Game 3), and completing the rest of the games on their own. Four frequency bands of brain waves were recorded for each game, and a total of 24 sets of EEG data were recorded for the two groups.

4.3 Data Analysis

We used Discrete Fourier Transform (DFT) to calculate the power spectral density (PSD) of each participant's EEG waves, to get to its overall power and to assess whether there is a difference between the four waves of the participants in the high age group and the low age group, and we processed the PSD values we obtained, and according to the third section of this paper, it can be seen that the difference analysis algorithm based on the Serious Game Test Model we designed was used for the PSD values. processing. When the data did not satisfy the assumption of normality or variance chi-square, appropriate nonparametric tests were used or the data were transformed and obtained to get the D-statistic, and finally its corresponding p-value was reported to determine whether there was a significant difference in power among different groups or under different conditions. The level of significance was set at 0.05. Based on the data obtained after processing, we summarized the calculated results. Game 1 refers to the game of Straw Boat Borrowing Arrows, and Game2

refers to the game of Thorng Khor Assassinate Qin. Groups H and L represent the high age group and low age group, respectively.

Table 1. Table analyzing the differences in power spectral density (PSD) values of the four brainwaves in Game1

Waves	Source of Variation	SS	df	MS	D-statistic	P-Value
α	Between Groups (Sa)	15.24	1	15.24	5.28	0.035*
	Within Groups (Se)	34.16	18	-	-	-
	Total	49.80	19	-	-	-
θ	Between Groups	22.40	1	22.40	12.89	0.003*
	Within Groups	31.20	18	1.73	-	-
	Total	53.60	19	-	-	-
β	Between Groups	22.50	1	22.50	12.00	0.004*
	Within Groups	33.75	18	1.875	-	-
	Total	56.25	19	-	-	-
δ	Between Groups	1.20	1	1.20	0.95	0.10
	Within Groups	25.30	18	0.66	-	-
	Total	26.50	19	-	-	-

As can be seen from Table 1, the p-value of α-wave obtained in Game1 is 0.035, which means that the difference of α-wave is significant ($p<0.05$) on groups H and L. The p-value of θ-wave is 0.003, which is more significant compared to α-wave's 0.035, because the smaller p-value indicates that the result is less likely to be related to chance. β-wave has a p-value of 0.004, which is not very much different from θ-wave, and β-wave can be surely not happened by chance. There is not much difference, and theta and beta waves can be sure that they did not occur by chance. The p-value of δ-wave is 0.10, so the difference between δ-wave of group H and group L is not considered significant ($p>0.05$), which means that the change of δ-wave in Game may be caused by random fluctuation only.

The p-value of α-wave obtained by group H and group L in Game2 is 0.033, which means that the difference between group H and group L in α-wave is significant ($P<0.05$). The p-value of θ-wave is 0.004, which is more significant compared to 0.035 for α-wave. As mentioned in the analysis of Table1, the smaller p-value means that the result is less likely to be related to chance. The p-value of β-wave is 0.007, which is similar to the P-value of θ-wave, θ-wave and β-wave can be sure that they did not happen by chance. Unlike the δ-wave, the P-value of δ-wave is 0.10, so the difference between Group H and Group L in δ-wave is considered insignificant ($P>0.05$), which means that the change of δ-wave in the Game may be caused by random fluctuation only.

The p-values of α, β and δ waves obtained in Game3 for Groups H and L were 0.06, 0.09 and 0.14 respectively, which were all >0.05,which means that the difference between Groups H and L on these three waves was insignificant, while the p-value of θ wave was higher compared to that in Game1 and Game2, which indicates that although the difference between Groups H and L on θ wave is still significant, it is still not significant in comparison to the differences between Groups Game1 vs.

Game2, the level of significance has decreased. By comparing with the data obtained in Game1 and Game2, it can be seen that external guidance has a greater impact on the brain waves of children in Group L during play, especially in the α and β waves, and these results suggest that guidance can help children in the younger age group to focus their attention and improve their thinking ability, especially when it comes to creative and problem-solving tasks, and the younger age group improved their cognitive load level under guidance. Theta waves, on the other hand, were related to children's memory processes and emotional responses, which were not significantly altered by external guidance.

5 Discussion

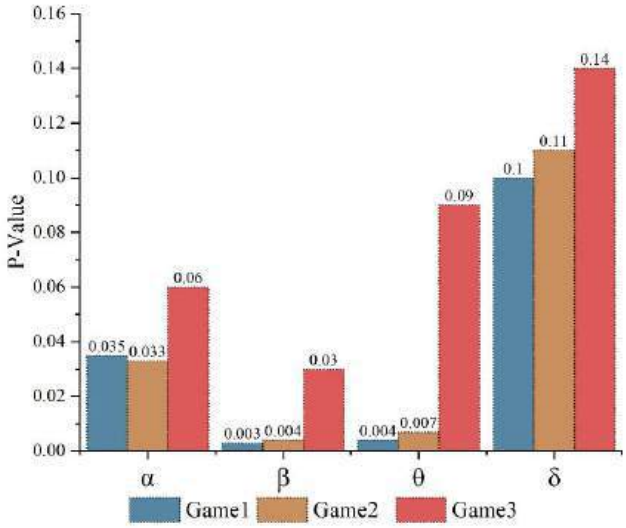


Fig. 3. Histogram of P-values for four waves in three games

Based on the 12 p-value results obtained from the experiments, we plotted a histogram of p-values (see Fig.3), according to the histogram, we compared the p-values of the same waves in Game1 and Game2 two by two, the p-values of the same waves did not have much difference, they were all controlled within 0.1, while the p-values of the different waves differed significantly, which was due to the fact that the magnitude of the p-values received the influence of a variety of factors, and that different EEG waves were associated with different types of cognitive and neurophysiological processes, and each wave has different brain region activities, differences in electrode placement, signal processing and analysis methodology may also affect the p-value of a particular band, many reasons may cause this difference in p-values between different waves. The difference in p-values between group H and group L in the same wave is not large, this may be due to the similarity in effect sizes, sample sizes, and statistical power, and if the p-values of the two age groups have

similar p-values, this means that the effect of age on that particular variable may be statistically similar when controlling for other variables. p-values are also affected by sample size. p-values for Game3 species differed significantly from those of Game1 and Game2 species, where the α, β, and δ waves were greater than 0.05, and the cause of this was related to the fact that Group L received bootstrapping in Game3 species that appropriate guidance mechanisms have a positive impact on younger children, especially in terms of enhancing their immersion and learning in games.

The aim of this study was to investigate the effects of serious games on brain waves (theta, alpha, beta and delta waves) in participants of different age groups. To this end, a well-designed serious game-based test model was used to analyze the effect of the game on EEG activity, and further difference-in-difference analyses were conducted to reveal statistically significant key findings. After taking a deeper look at the EEG waveforms recorded during gameplay, we noticed significant age-group differences.

For participants in the lower age group (Group L), the activity and amplitude of theta and alpha waves were relatively high, indicating that this group was more prone to deep attentional focus and reflective thinking during gameplay. This finding inspires us to emphasize the importance of adding elements that can capture attention and stimulate thinking when designing serious games for the L group. By doing so, we can expect to maintain and increase the interest and effectiveness of children in this age group. In contrast, participants in the older age group (Group H) showed higher beta-wave activity, reflecting a higher level of alertness, concentration, and information-processing abilities during play compared to younger children. Therefore, more complex challenges and tasks should be incorporated into the design of serious games for Group H children to promote their detailed attention and higher-order cognitive skill development.

In particular, when applied to children in the younger age group, our study also found that when using Game3, a game for intervention experiments, alpha and beta waves showed significant changes after appropriate instruction and guidance. This result suggests that the effectiveness of serious game design lies not only in the content of the game itself, but also in the accompanying guidance methods. Proper guidance plays an important role in enhancing children's immersion in the game in the younger age group, especially in expanding their cognitive ability, concentration, and immersion experience showing significant positive effects.

Taken together, our study highlights the unique role that serious games can play in the cognitive development of participants of different ages, and provides important insights into how to optimize game design and assistive guidance for specific age groups to promote effective learning and development.

6 Conclusion

In this study, we conceived and executed an educational experiment to investigate differences in immersion levels among primary school-aged children engaged in serious games. By creating a difference analysis algorithm based on a serious game

test model, we delved into neurophysiological metrics associated with immersion, specifically power spectral density (PSD).

The results reveal significant differences in the neurophysiological facets of immersion among children in different age brackets. Younger children demonstrated more active relaxation and creative thinking patterns, as reflected in their brainwave activity. Conversely, older children exhibited increased focus and increased alertness. These findings not only deepen our understanding of the differences in immersion induced by serious games in virtual reality settings but also unveil a connection between cognitive developmental stages and electrophysiological indicators. Well-guided interventions can substantially enhance immersion in games for younger children.

Future studies may expand their scope to consist of a broader range of populations and game genres, building upon the findings of this project to verify and enrich our findings. Nonetheless, it is essential to acknowledge the limitations of this study, such as the relatively small sample size, which may constrain the generalizability of the findings. As the sample size grows and age stratification becomes more refined, we anticipate further validation and expansion of these findings. Additionally, the effects of various aspects of serious game design, such as difficulty level, storyline, or interactivity, on immersion and electrophysiological responses warrants further exploration.

In conclusion, research has demonstrated the significant potential of virtual reality technology and serious games in children's education. The creation and implementation of customized pedagogical tools tailored to the cognitive attributes of children in different age groups are necessary for realizing each learner's optimal learning potential. As technology continues to advance, we eagerly anticipate the development of more precise and captivating educational games capable of effectively stimulating children's interest in learning and unlocking their latent abilities. Clearly defining the target audience for serious games can specifically enhance their education.

Acknowledgments. This work is supported in part by the Research Project of Humanities and Social Sciences of the Ministy of Education with grant No. 24YJAZH075, International Cooperation Project of Henan Province with grant No.252102520012, the Research Project of Humanities and Social Sciences of Henan Province with grant No. 2025-ZZJH-370, the Research Project of Intangible Cultural Heritage of Henan Province with grant No. 24HNFY-LX149, the Postgraduate Education Reform and Quality Improvement Project of Henan Province with grant No. YJS2025AL39.

Disclosure of Interests. We declare that we have no fnancial and personal conficts of interests with other people or other organizations that may inappropriately infuence our work. There are no professional or personal conficts of interests of any nature or any kind in any product, service and/or company that could be construed as influencing the position presented in, or the review in, the manuscript entitled.

Ethics approval All procedures performed in studies involving human partici-pants were in accordance with the ethical standards of the institutional and/or national research committee

and with the 1964 Hel-sinki declaration and its later amendments or comparable ethical standards. The IRB approval number is PMSM-20210227.

Availability of data and materials All data generated or analysed during this study are included in this published article (and its supplementary information fles). Requests for material should be made to the corresponding authors.

References

1. Michaelis, J.E., & Mutlu, B.: Supporting Interest in Science Learning with a Social Robot. Proceedings of the 18th ACM International Conference on Interaction Design and Children. pp. 71-82. (2019)
2. Bergin, D.A.: Social Influences on Interest. Educational Psychologist, 51, 22-7 (2016).
3. Ahmadov, T., Karimov, A., Durst, S., Saarela, M., Gerstlberger, W., Wahl, M. F., & Karkkainen, T.: A two-phase systematic literature review on the use of serious games for sustainable environmental education. Interactive Learning Environments, 33(3), 1945–1966 (2024)
4. Žilak, M., & Car, Ž.: A Framework for Improving Accessibility of Serious Games in Handheld Augmented Reality Based on User Interaction Data. Applied Sciences, 15(4), 2161 (2025)
5. Villada Castillo, J. F., Bohorquez Santiago, L., & Martínez García, S.: Optimization of Physics Learning Through Immersive Virtual Reality: A Study on the Efficacy of Serious Games. Applied Sciences, 15(6), 3405 (2025)
6. Gundersen, S. W., & Lampropoulos, G.: Using Serious Games and Digital Games to Improve Students' Computational Thinking and Programming Skills in K-12 Education: A Systematic Literature Review. Technologies, 13(3), 113 (2025)
7. Brown, E., & Cairns, P.A.: A grounded investigation of game immersion. CHI EA '04, pp.1297-1300 (2004)
8. Pange, J., Lekka, A., & Katsigianni, S. Serious Games and Motivation.: Conference on Interactive Mobile Communication Technologies and Learning, pp.240-246 (2017)
9. Barclay, P.A., & Bowers, C.A.: Associations of Subjective Immersion, Immersion Subfactors, and Learning Outcomes in the Revised Game Engagement Model. Int. J. Game Based Learn., 8, 41-51 (2018)
10. Hsiao, H.: A Brief Review of Digital Games and Learning. 2007 First IEEE International Workshop on Digital Game and Intelligent Toy Enhanced Learning (DIGITEL'07), pp.124-129 (2007)
11. Goli, A., Teymournia, F., Naemabadi, M., & Garmaroodi, A.A.: Architectural design game: A serious game approach to promote teaching and learning using multimodal interfaces. Education and Information Technologies, 27, 11467 – 11498 (2022)
12. Cone, B.D., Thompson, M.F., Irvine, C.E., & Nguyen, T.D.: Cyber Security Training and Awareness Through Game Play. IFIP International Information Security Conference, pp.432-436 (2006)
13. Kroustalli, C., & **nogalos, S.: Studying the effects of teaching programming to lower secondary school students with a serious game: a case study with Python and CodeCombat. Education and Information Technologies, 26(5), 6069-6095 (2021)
14. Gros, B.: "Digital games in education: The design of games-based learning environments" , Journal of Research on Technology in Education, , 40(1), 23-38 (2007)

15. VanDeventer, S. S., & White, J. A.: Expert behavior in children's video game play. Simulation & Gaming, 33(1), 28-48 (2002)
16. Cornejo, R., Martínez, F., Álvarez, V. C., Barraza, C., Cibrian, F. L., Martínez-García, A. I., & Tentori, M.: Serious games for basic learning mechanisms: reinforcing Mexican children's gross motor skills and attention. Personal and Ubiquitous Computing, 25, 375-390 (2021)
17. Westera, W., Prada, R., Mascarenhas, S., Santos, P. A., Dias, J., Guimarães, M., ... & Ruseti, S.: Artificial intelligence moving serious gaming: Presenting reusable game AI components. Education and Information Technologies, 25(1), 351-380 (2020)
18. Lindberg, R.S., & Laine, T.H.: Formative evaluation of an adaptive game for engaging learners of programming concepts in K-12. Int. J. Serious Games, 5(2), 3-24 (2018)
19. Sajjadi, P., Broeckhoven, F.V., & Troyer, O.D.: Dynamically Adaptive Educational Games: A New Perspective. International Conference on Serious Games pp. 71-76 (2014)
20. Cheng, M.T., She, H., & Annetta, L.A.: Game immersion experience: its hierarchical structure and impact on game-based science learning. J. Comput. Assist. Learn., 31(3), 232-253. (2015)
21. Barclay, P. A., & Bowers, C.: Associations of subjective immersion, immersion subfactors, and learning outcomes in the revised game engagement model. International Journal of Game-Based Learning, 8(1), 41-51 (2018)
22. Antonaci, A., Klemke, R., & Specht, M.M.: Towards Design Patterns for Augmented Reality Serious Games. International Conference on Mobile and Contextual Learning pp.273-282 (2015)
23. Greipl, S., Moeller, K., Kiili, K., & Ninaus, M.: Lifelong learning with a digital math game: Performance and basic experience differences across age. In Games and Learning Alliance: 8th International Conference, GALA 2019, Athens, Greece, November 27–29, 2019, Proceedings 8 (pp. 301-311). Springer International Publishing (2019)
24. Wan, B., Huang, W., Bai, L., & Guo, J.: Using Support Vector Machine on EEG Signals for College Students' Immersive Learning Evaluation. In 2021 7th International Conference of the Immersive Learning Research Network (iLRN), pp. 1-5. IEEE (2021)
25. Ray, W. J., & Cole, H. W.: EEG alpha activity reflects attentional demands, and beta activity reflects emotional and cognitive processes. Science, 228(4700), 750-752 (1985)
26. Laufs, H., Kleinschmidt, A., Beyerle, A., Eger, E., Salek-Haddadi, A., Preibisch, C., & Krakow, K.: EEG-correlated fMRI of human alpha activity. Neuroimage, 19(4), 1463-1476 (2003)
27. Škola, F., Rizvić, S., Cozza, M., Barbieri, L., Bruno, F., Skarlatos, D., & Liarokapis, F.: Virtual reality with 360-video storytelling in cultural heritage: Study of presence, engagement, and immersion. Sensors, 20(20), 5851 (2020)
28. Tauscher, J. P., Schottky, F. W., Grogorick, S., Bittner, P. M., Mustafa, M., & Magnor, M.: Immersive EEG: evaluating electroencephalography in virtual reality. In 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp. 1794-1800. IEEE (2019)
29. Shu, Y., Huang, Y., Chang, S., & Chen, M.: Do virtual reality head-mounted displays make a difference? A comparison of presence and self-efficacy between head-mounted displays and desktop computer-facilitated virtual environments. Virtual Reality, 23, 437-446 (2018)
30. Barrett, A. J., Pack, A., & Quaid, E. D.: Understanding learners' acceptance of high-immersion virtual reality systems: Insights from confirmatory and exploratory PLS-SEM analyses. Computers & Education, 169, 104214 (2021)

31. Huang, W., Roscoe, R. D., Johnson-Glenberg, M. C., & Craig, S. D.: Motivation, engagement, and performance across multiple virtual reality sessions and levels of immersion. Journal of Computer Assisted Learning, 37(3), 745-758 (2 021)
32. Acharya, R.U., Faust, O., Alvin, A.P., Sree, S.V., Molinari, F., Saba, L., Nicolaides, A.N., & Suri, J.S.: Symptomatic vs. Asymptomatic Plaque Classification in Carotid Ultrasound. Journal of Medical Systems, 36, 1861-1871 (2012)

Unsupervised Salient Object Detection with Pseudo-Labels Refinement

Yanfeng Zheng[1], Pengjie Wang[1,2], Hao Liu[1], and Xiaosong Yang[2⋆]

[1] Dalian Minzu University, Dalian 116650, China
Zhengyanfeng1998@163.com
pengjiewang@gmail.com
202412054063@stu.dlnu.edu.cn
[2] Bournemouth University, Fern Barrow, Poole, Dorset, BH12 5BB, United Kingdom
xyang@bournemouth.ac.uk

Abstract. In Salient Object Detection(SOD), most methods rely on manually annotated labels, which are costly. As a result, unsupervised methods have gained significant attention. Existing methods often generate noisy pseudo-labels using traditional techniques, which can affect model performance. To address this, we propose an unsupervised method for RGB image salient object detection that generates high-quality pseudo-labels without manual annotation and uses them to train the detection model.The method generates initial pseudo-labels and improves their quality by introducing contrastive learning pre-trained weights and a pseudo-label self-updating strategy. Additionally, we design a detection network with a Multi-Feature Aggregation (MFA) module and a Context Feature Interaction (CFI) module to enhance the model's ability to detect salient objects in complex scenarios. The model we proposed, trained with our pseudo-labels, shows significant improvement on USOD and achieves excellent scores on public benchmarks.

Keywords: Unsupervised · Salient Object Detection · Contrastive Learning · Pseudo-Labels.

1 Introduction

The development of deep learning has significantly advanced salient object detection, with fully-supervised methods achieving notable breakthroughs. However, these methods are highly dependent on large-scale, accurately labeled data. To reduce the annotation burden, weakly-supervised methods have emerged, such as class labels [1] text descriptions [2],bounding boxes [3], scribbles [4] and point annotations [5]. Despite progress, human annotation is still required. Unsupervised methods aim to eliminate the need for human annotations altogether, offering better applicability in real-world scenarios where labeled data is scarce. A key challenge for unsupervised methods is generating high-quality pseudo-labels through image modeling, which is essential for training effective models.

⋆ *Corresponding author

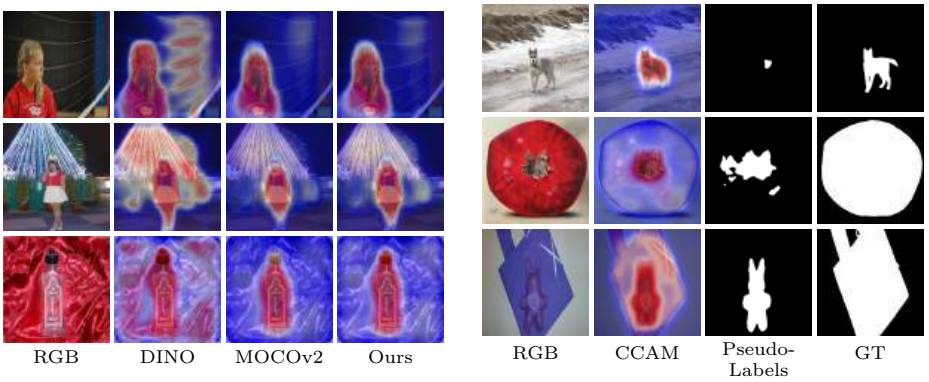


Fig. 1. (a) Visualisation of class-agnostic activation maps for different pre-trained weights. (b) Incorrect pseudo-labeling results.

Before the rise of deep learning, unsupervised methods mainly relied on handcrafted features like color contrast to identify salient regions, but these methods struggled in complex scenes. Today, most unsupervised methods generate initial pseudo-labels using traditional techniques and refine them with various strategies. However, traditional methods often produce low-quality pseudo-labels, limiting detection performance. Researchers are exploring advanced algorithms to improve pseudo-label accuracy and overall detection. Few methods use deep learning for pseudo-label generation, but Zhou et al. [6] showed that pre-trained weights from convtrastive learning can provide supervision for salient object detection models, yielding impressive results. One such method, CCAM [7], uses unsupervised contrastive learning to identify foreground regions by contrasting foreground and background in different images. As shown in Figure 1, CCAM trained with MOCOv2 [8] weights achieves good foreground localization but incomplete coverage, while CCAM trained with DINO weights [9] provides full coverage but with redundancy. These issues affect the quality of the final pseudo-labels.

In generating category-agnostic activation maps and refining them with a dense conditional random field (DCRF) to produce pseudo-labels, several challenges arise, as shown in Figure 1. While activation maps highlight target regions, they often lack precise edges, and complex scenes present further refinement difficulties. Additionally, some activation regions may not be suitable for salient object detection, leading to inaccurate pseudo-labels. To address these issues, this paper proposes a two-stage model for salient object detection. The first stage generates pseudo-labels in two steps: enhancing the original CCAM using offline distillation for the initial pseudo-label network, and refining the labels with a self-updating strategy. The second stage focuses on salient object detection, where the model is primarily supervised by the generated pseudo-labels. Key components of this model include: 1) a multi-feature aggregation module

to enhance high-level features, and 2) a context feature interaction module for improved feature fusion, boosting detection performance.

Our main contributions can be summarized as follows:

(1) This work introduces an updated pseudo-label generation method, leveraging different pre-trained weights for complementary learning and a self-updating strategy to improve label quality.

(2) A salient object detection network is designed to boost detection performance, incorporating a multi-feature aggregation module and a context feature interaction module.

(3) Experiments on four common RGB image saliency detection datasets demonstrate that the proposed method performs comparably to current weakly-supervised and unsupervised approaches.

2 Related work

2.1 Fully-Supervised Method Salient Object Detection

The majority of Salient Object Detection (SOD) methods are rely on extensive pixel-level manual annotations as the foundation for training and optimization. Qin et al. [10] proposed the BASNet method, which incorporates boundary-aware mechanisms to enhance the accuracy of salient object detection by focusing on the boundaries of objects. Liu et al. [11] proposed a feature aggregation module structure based on the U-net structure, combining coarse-level and high-level information. Pang et al. [24] proposed a multi-scale interactive network that uses multi-scale features and interactive mechanisms to improve the accuracy of salient object detection. Xu et al. [13] proposed PA-KRN, a progressive architecture for salient object detection that first locates objects globally using a coarse module, then segments them locally with a fine module, and uses an attention-based sampler to highlight salient regions. Liang et al. [14] proposed ExPert, a parameter-efficient fine-tuning method for salient object detection that uses adapters and injectors in a frozen transformer encoder to incorporate external prompt features, achieving superior performance with fewer parameters.

2.2 Weakly-Supervised Method Salient Object Detection

The prevailing state-of-the-art techniques for salient object detection are heavily dependent on extensive datasets that require precise pixel-level manual annotations. The creation of such annotations is both time-consuming and labor-intensive. Consequently, weakly-supervised approaches are emerging as a prominent and increasingly favored research trajectory. Piao et al. [15] employed an iterative calibration strategy to mitigate the pseudo-labeling error within the network. Zhang et al. [16] conducted supervised training by annotating simple pairs of images with foreground and background labels. Piao et al. [17] introduced a multiple pseudo-label fusion framework that leverages richer information from multiple labels to diminish the impact of the algorithmic process. Gao et al. [18]

presented a point-supervised approach that initially acquires pseudo-labels via an adaptive masking algorithm and subsequently generates the final prediction saliency maps through a Transformer-based network.

2.3 Unsupervised Method Salient Object Detection

In the field of salient object detection, weakly-supervised methods have played a significant role, but unsupervised methods have also garnered considerable attention. Unsupervised methods aim to detect salient objects without any explicit annotations. Nguyen et al. [19] proposed the DeepUSPS method, which uses self-supervision to leverage the input image itself as a natural supervisory signal for robust unsupervised saliency prediction. Yan et al. [20] introduced an uncertainty-aware pseudo-label learning approach for unsupervised domain adaptation in salient object detection, enabling the model to adapt to the target domain without labeled data in that domain. Wang et al. [21] proposed a method for deep unsupervised saliency detection that mines multi-source uncertainty to select reliable labels from multiple noisy labels, thereby improving the performance of unsupervised saliency detection. Zhou et al. [6] introduced a method called "Activation to Saliency", which forms high-quality labels for unsupervised salient object detection by leveraging activation information, leading to better detection results. Zhou et al. [22] proposed a texture-guided saliency distilling method by matching textures around the predicted boundaries for unsupervised salient object detection.

3 Method

The unsupervised saliency object detection process discussed in this paper mainly consists of two key stages: the first is the pseudo-label generation stage, where pseudo-labels are generated based on RGB images; the second is the saliency object detection stage, which differs from fully-supervised methods in that it uses the pseudo-labels generated in the first stage for learning and supervision. In this section, we will first describe the method for generating pseudo-labels, and then introduce the two core modules that constitute the saliency object detection network, namely the Multi-Feature Aggregation module (MFA) and the Contextual Feature Interaction Module (CFI).

3.1 Pseudo-label generation model

This study proposes a novel method for generating pseudo-labels using class-agnostic activation maps, which automatically identify and locate salient objects. Instead of directly using the CCAM method, the network is enhanced with different pre-trained weights. A CCAM model trained with DINO pre-trained weights serves as an auxiliary supervision signal, providing additional guidance to improve training and combine the strengths of both weight sets.

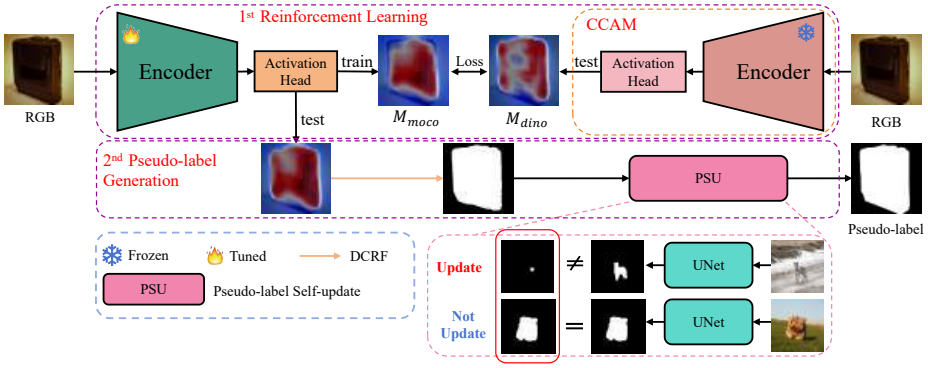


Fig. 2. Pseudo-label generation method structure

As shown in the upper part of Figure 2, in the specific implementation, Resnet-50 is used as the encoder of the backbone network. An RGB image is input, and after being processed by the encoder of the backbone network, four sets of feature maps F_1, F_2, F_3, and F_4 are obtained. This process can be represented as:

$$F_1, F_2, F_3, F_4 = Encoder(I_m) \tag{1}$$

Here, I_m represents the input RGB image, and Encoder represents the encoder. Then, the feature maps F_3 and F_4 are concatenated along the channel dimension and then processed through the CBS operation to generate the class-agnostic activation map M_{moco}, This process can be represented as:

$$M_{moco} = CBS(Contact(F_3, F_4)) \tag{2}$$

Here, $Concat()$ denotes the concatenation operation along the channel dimension, and CBS represents a sequence of operations including a 3×3 convolution, BatchNorm, and a Sigmoid activation function. Additionally, based on the aforementioned process, the encoder is pre-trained using DINO pre-trained weights to generate a class-agnostic activation map represented as M_{dino}.

$$\mathcal{L} = \mathcal{L}_{\text{POS}} + \mathcal{L}_{\text{NEG}} + \alpha\mathcal{L}_{\text{SSIM}} + \beta\mathcal{L}_{\text{IoU}} \tag{3}$$

Here, $\mathcal{L}_{\text{POS}}$ and $\mathcal{L}_{\text{NEG}}$ are the original CCAM losses, $\mathcal{L}_{\text{SSIM}}$ is the structural similarity loss, and $\mathcal{L}_{\text{IOU}}$ is the intersection over union loss. The values of α and β are set to 0.2.

After generating the final class-agnostic activation maps using the aforementioned strategy, Dense Conditional Random Fields (DCRF) are further employed to process these activation maps to generate the initial pseudo-labels Y_{PL}. This process aims to refine the saliency maps from the original activation maps, providing more accurate labels for subsequent training. However, although DCRF can improve the quality of the labels to some extent, the pseudo-labels still have imperfections in detail, as shown in the first and second columns of the third row in Figure 1. Due to the characteristics of the class-agnostic activation maps,

some activated regions may not be entirely suitable for the task of salient object detection, as shown in the third and fourth columns of the third row in Figure 1. These incomplete or incorrect refinements, if used as the basis for long-term network training, may lead the model to learn these inaccurate pieces of information, ultimately affecting the detection performance of the network. Despite the

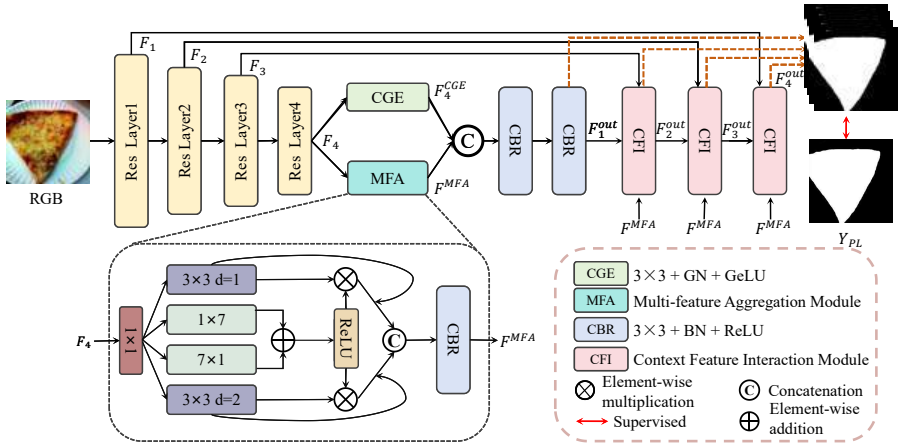


Fig. 3. The structure of the salient object detection network

potential inaccuracies in the pseudo-labels, network training remains an iterative learning and optimization process. Even with imperfect labels, they still guide the salient object detection network towards the correct targets, providing a generally valid learning direction. This demonstrates that the network can learn effective saliency information by capturing statistical patterns in large datasets, even with imprecise labels. In the early stages of training, the network is highly sensitive to the saliency information in the pseudo-labels, highlighting the importance of effective pseudo-label updating strategies. A well-designed updating strategy enhances the network's ability to capture saliency features, improving detection performance. Based on this, we propose a pseudo-label self-updating algorithm, as shown in the lower part of Figure 2. Specifically, the generated pseudo-labels Y_{PL} are used to train a simple U-shaped network, and the saliency map $Y_{PL}^{'}$ produced by the network is used to update the pseudo-labels. In the early stages, the model can more accurately identify and correct errors in the pseudo-labels, and iteratively updating them improves both their accuracy and detail, ultimately enhancing the detection performance.

In this algorithm, the pseudo-labels are self-updated using different evaluation criteria at different training stages to improve the model's performance. Specifically, in the 2nd to 5th rounds of training, the algorithm uses the intersection over union (IoU) to measure the similarity between the model's current predictions and the previous pseudo-labels.If the result is below the threshold, the pseudo-labels are updated using the current model predictions. In the later

stages of training, the pseudo-labels are updated using the Structure Similarity Index Measure (SSIM) [34] as the update criterion.

Here, the threshold is initially set to 0.9 for each evaluation criterion, and starting from the second epoch it is continuously updated during training, increasing by 0.1 each epoch over a total of 10 epochs. By dynamically adjusting the update strategy during training, the pseudo-labels are continuously refined, thereby enhancing the model's understanding of the data and the accuracy of its predictions.

3.2 Unsupervised Salient Object Detection with Pseudo-labels

To better enhance the performance of salient object detection, this paper designs a salient object detection model that uses Resnet-50 as the backbone network for feature extraction. An input RGB image is processed through the backbone network to obtain four features, namely F_1, F_2, F_3, and F_4, which are used as inputs for the multi-feature aggregation module and the context feature interaction module. The overall architecture is shown in Figure 3.

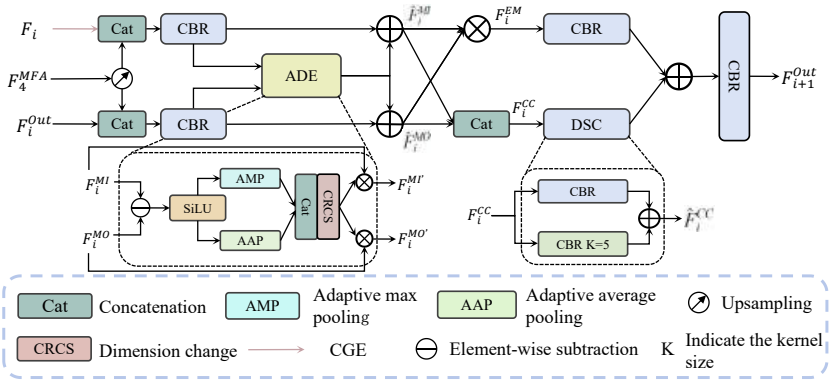


Fig. 4. Contextual Feature Interaction Module (CFI)

Multi-Feature Aggregation Module In deep learning tasks, the shallow layers of a network extract low-level features, while higher convolutional layers extract more advanced features. Among these, high-level semantic features are crucial as they provide a deep and abstract understanding of the image content. The abstract nature of these features enables them to effectively capture complex concepts and entities within the image, ensuring robustness against variations. By enhancing high-level semantic features, the model can more accurately understand and represent complex structures and abstract concepts within the image. Chen et al. [27] used dilated convolutions to expand the receptive field of convolutional layers, significantly improving the model's ability to recognize objects of different sizes without increasing the number of parameters or computational

burden. To this end, this paper designs a multi-feature aggregation module that primarily enhances the high-level feature F_4 from the encoder. By employing convolutional kernels of various sizes and shapes, the module enhances the feature representation and adapts to the processing needs of objects of different shapes. Specifically, as shown in the MFA (Multi-feature Aggregation) module in Figure 3, the input is F_4. First, a 1×1 convolution is applied to reduce the dimensionality of the feature, resulting in $F_4^{'}$. $F_4^{'}$ is then processed through 3×3 convolution operations with different dilation rates to obtain the features $\tilde{F}_4$ and$\bar{F}_4$. The process can be represented as:

$$\begin{aligned} \tilde{F}_4 &= \mathrm{Conv}_{d=1}(F_4') \\ \bar{F}_4 &= \mathrm{Conv}_{d=2}(F_4') \end{aligned} \tag{4}$$

Here, Conv denotes a convolution with a 3×3 kernel, and d represents the dilation rate. By combining vertical and horizontal convolution kernels, the model can more comprehensively capture spatial information in the image. Compared to using traditional 3×3 and 7×7 convolution kernels, this method not only reduces the number of parameters and the risk of overfitting but also increases the model's processing speed and efficiency. For this reason, $F_4^{'}$ is also processed through convolution kernels in different directions to obtain spatial information in the image and then passed through a ReLU layer to obtain F_{HW}. The process can be represented as:

$$F_{HW} = \mathrm{ReLU}(\mathrm{Conv}_H(F_4') \oplus \mathrm{Conv}_W(F_4')) \tag{5}$$

Here, $Conv_H$ denotes a vertical convolution with a 7×1 kernel, and $Conv_W$ denotes a horizontal convolution with a 1×7 kernel. The symbol $\otimes$ represents element-wise addition. To better integrate the features from dilated convolutions and the spatially enhanced features, the feature map F_{HW} is element-wise multiplied with the dilated features $\tilde{F}_4$ and $\bar{F}_4$ of different dilation rates. Additionally, skip connections are applied to each set of features to fuse the original features. This approach not only enhances the spatial representation but also maintains the integrity of the original features, thereby providing the network with a richer and more effective feature representation.

$$\begin{aligned} \tilde{F}_4 &= \tilde{F}_4 \copyright (F_{HW} \otimes \tilde{F}_4) \\ \bar{F}_4 &= \bar{F}_4 \copyright (F_{HW} \otimes \bar{F}_4) \end{aligned} \tag{6}$$

Here, $\oplus$ denotes element-wise multiplication. Finally,$\tilde{F}_4$ and $\bar{F}_4$ are concatenated and then passed through a CBR to obtain the feature F_{MFA}. The process can be represented as: Through the aforementioned operations, convolutional kernels of different shapes and sizes are effectively integrated, thereby significantly enhancing the feature representation capabilities. By expanding the receptive field, this method enables the network to learn richer spatial attributes, thereby deeply exploring and utilizing the complexity and diversity of image content. This enhances the high-level feature F_4 and provides richer and more effective input features for subsequent modules.

Context Feature Interaction Module In salient object detection, the U-shaped structure is commonly used for its strong performance. However, as high-level features pass upwards in this structure, their information density decreases, impacting detection capability [11]. To address this, we propose a Context Feature Interaction Module that enhances feature interaction across levels, mitigating the dilution of high-level features during transmission.

As shown in Figure 4, the inputs to this module are F_{MFA}, F_i^{Out}, and F_i, which originate from different stages of the model and each contain unique information and data representations. First, F_{FMA} is concatenated with F_i^{Out} and F_i respectively. Then, these concatenated features are processed through two separate CBRs to obtain two new features F_i^{MI} and F_i^{MO}. These features are then fed into the Adaptive Difference Enhancement Module (ADE).

The primary function of the ADE module is to calculate the differences between the two input features and process these difference features using the SiLU function to highlight important information and suppress less important information. Subsequently, the ADE module further processes these difference features through adaptive average pooling and adaptive max pooling operations. These two types of pooling operations extract features from different perspectives, and combining the pooled features helps to integrate their respective advantages. By applying these combined features to the original input features through element-wise multiplication, the expressive power of the input features is further enhanced. Additionally, skip connections are introduced to prevent information loss during the weighting process, resulting in $\hat{F}_i^{MI}$ and $\hat{F}_i^{MO}$. The process is as follows:

$$\begin{aligned} F_i^{MI'}, F_i^{MO'} &= ADE(F_i^{MI}, F_i^{MO}) \\ \hat{F}_i^{MI} &= F_i^{MI'} + F_i^{MI} \\ \hat{F}_i^{MO} &= F_i^{MO'} + F_i^{MO} \\ F_i^{CC} &= \mathrm{Cat}(\hat{F}_i^{MI}, \hat{F}_i^{MO}) \end{aligned} \tag{7}$$

In the feature interaction operation, $\hat{F}_i^{MI}$ and $\hat{F}_i^{MO}$ are element-wise multiplied to generate F_i^{EM}, which helps to capture and enhance the interactions and dependencies between the two features.

$$F_i^{EM} = \hat{F}_i^{MI} \otimes \hat{F}_i^{MO} \tag{8}$$

To enhance the representation capability of the feature $\hat{F}_i^{CC}$, a multi-scale convolutional kernel strategy is employed to capture different scale information from the input features. Specifically, convolutional kernels of different sizes are applied to $\hat{F}_i^{CC}$ to extract features at different scales, and these features are then element-wise added to obtain F_i^{CC}. The process can be represented as:

$$\hat{F}_i^{CC} = CBR(F_i^{CC}) + CBR_{k=5}(F_i^{CC}) \tag{9}$$

By integrating features from different scales, the expressiveness and adaptability of the features are further enhanced. Finally, to combine multiple feature representations, $\hat{F}_i^{CC}$ andF_i^{EM} are element-wise added and then processed through a

CBR operation to obtain the final output feature F_{i+1}^{Out} of the Context Feature Interaction Module. The process can be represented as:

$$F_{i+1}^{Out} = CBR(\hat{F}_i^{CC} + F_i^{EM}) \tag{10}$$

This paper replaces the traditional U-shaped structure's decoder with the Context Feature Interaction Module, which more effectively integrates feature information across different levels, particularly during upsampling and resolution restoration. This module combines deep semantic information with shallow detail, enhancing the model's ability to capture target details and improving overall feature representation. As a result, the model better incorporates both contextual and local information during decoding, boosting performance.

3.3 Loss Function

In this paper, a combined loss function is used for training, which includes the intersection over union loss ($\mathcal{L}_{IoU}$) and the local saliency coherence loss ($\mathcal{L}_{lsc}$) [25]. Additionally, this paper employs a deep supervision strategy, which introduces supervision signals at different network layers to further improve the model's performance. The formula for the total loss in this paper is as follows:

$$\mathcal{L} = \sum_{i=1}^{4} \left(\mathcal{L}_{IoU} \left(Y_i^{out}, Y_{pl} \right) + \mathcal{L}_{lsc} \right) \tag{11}$$

4 Experiments and results

4.1 Datasets

In the experiments of this paper, DUTS-TR [30], is used as the training dataset. The pixel-level pseudo-labels generated by the proposed method serve as supervision signals for network training. For testing, the method is evaluated on ECSSD [31], DUTS-TE [30], DUT-OMRON [32], and HKU-IS [33] datasets.

4.2 Experimental Details

Experiments were conducted on a NVIDIA GTX 3090 GPU using the PyTorch framework. The first stage's hyperparameters match those of CCAM, while the second stage uses a DINO pre-trained ResNet-50 as the backbone. Training images are resized to 256 $\times$ 256, with the Adam optimizer and a batch size of 32. The model trains for 15 epochs, starting with a learning rate of 1e-4, which decays by 10% every 5 epochs.

4.3 Evaluation Metrics

This paper employs three commonly used evaluation metrics in salient object detection, to assess the performance of different models. These include the F-measure (F_β) [28], Mean Absolute Error (MAE) [29], E-measure [26].

Table 1. Quantitative comparisons on four datasets

Method	Sup	DUTS-TE			HKU-IS			ECSSD			DUT-OMRON		
		$MAE \downarrow$	$E_m \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$	$E_m \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$	$E_m \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$	$E_m \uparrow$	$F_\beta \uparrow$
RBD [10]	T	0.162	0.664	0.428	0.176	0.716	0.54	0.206	0.705	0.577	0.165	0.654	0.416
BASNet [23]	F	0.048	0.884	0.791	0.032	0.946	0.895	0.037	0.921	0.88	0.056	0.869	0.756
MINet [24]	F	0.037	0.917	0.828	0.029	0.96	0.909	0.033	0.953	0.924	0.056	0.873	0.755
KRN [13]	F	0.034	0.926	0.851	0.028	0.959	0.916	0.036	0.92	0.922	0.049	0.889	0.783
WSSA [4]	W	0.062	0.869	0.742	0.047	0.932	0.86	0.059	0.917	0.870	0.068	0.845	0.703
MFNet [17]	W	0.079	0.832	0.692	0.058	0.919	0.839	0.084	0.880	0.844	0.098	0.784	0.621
SCWS [25]	W	0.049	0.907	0.823	0.038	0.943	0.896	0.049	0.931	0.900	0.060	0.870	0.758
USPS [19]	U	0.068	0.85	0.747	0.045	0.923	0.88	0.067	0.893	0.873	0.062	0.848	0.738
UDASOD [20]	U	0.05	0.897	0.795	0.035	0.947	0.883	0.043	0.94	0.895	0.059	0.849	0.733
UMNet [21]	U	0.067	0.863	0.752	0.041	0.939	0.889	0.064	0.904	0.879	0.063	0.860	0.743
A2S [6]	U	0.069	0.847	0.729	0.041	0.936	0.868	0.056	0.921	0.882	0.079	0.818	0.688
A2SV2 [22]	U	**0.047**	0.903	0.81	0.037	0.948	0.903	**0.044**	**0.940**	**0.917**	**0.061**	**0.864**	0.746
OURS	U	0.048	**0.905**	**0.822**	**0.033**	**0.953**	**0.915**	0.048	0.936	0.916	0.064	0.862	**0.752**

4.4 Comparison Experiments

This section compares the method proposed in this paper with fully-supervised, weakly-supervised, and unsupervised methods for salient object detection, including: RBD [10], BASNet [23], MINet [24], KRN [13], USPS [19], UDASOD [20], A2S [6], A2SV2 [22], MFNet [17], SCWS [35], UMNet [21], USPS [19] and WSSA [4]. The effectiveness of each method is evaluated by comparing the saliency maps they generate, either using the original code or directly provided by the authors. The comparisons aim to highlight the performance gap between unsupervised methods, which do not require manual annotations, and other supervised approaches. Additionally, the section emphasizes the performance of the proposed method, which operates without any manual annotations. All methods are evaluated using the same evaluation code to ensure fairness.

Quantitative Analysis The assessments are shown in Table 1. "Method" indicates the model name. "Sup" denotes the supervision method of the model, where "T" represents traditional methods, "F" indicates fully-supervised methods, "W" stands for weakly-supervised methods, and "U" signifies unsupervised methods. Results in bold font represent the best performance among unsupervised methods.

Qualitative Analysis As shown in Figure 5, compared with the current mainstream weakly-supervised and unsupervised methods, the method proposed in this paper demonstrates significant advantages on various types of images. Particularly in the first to second rows of images, the method in this paper performs excellently in detecting the salient object "door", almost accurately completing the segmentation of the region while maintaining the complete edges and detailed features of the "door". Compared with previous methods, they have deficiencies in detecting the details and edges of the "door". Furthermore, the method in this paper can accurately segment salient objects in complex scenes, as shown in the third to fourth rows. Additionally, it can precisely segment salient objects when

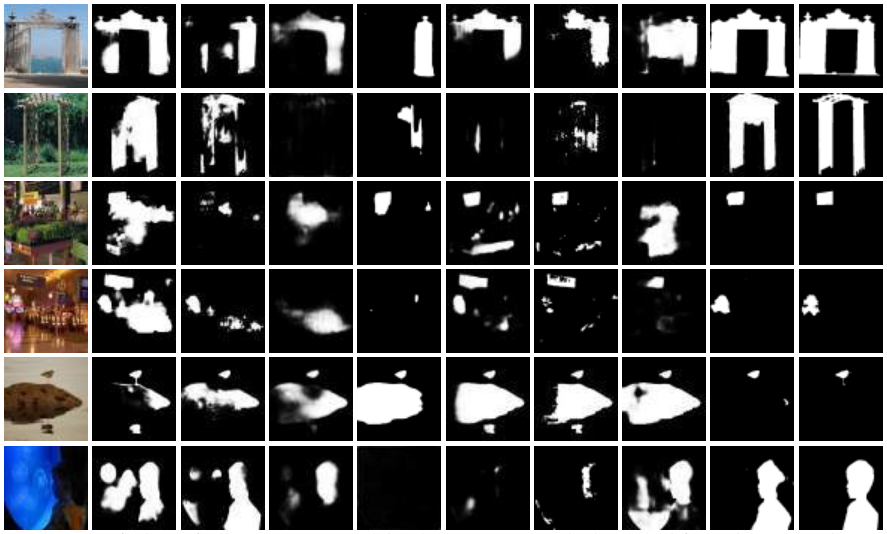


Fig. 5. Qualitative comparison of the methodology in this paper with other methods

they are small or when the input images have insufficient lighting. The above experimental results demonstrate the excellent performance of the method in this paper for salient object detection in complex tasks.

4.5 Ablation Studies

To evaluate the contributions of the various modules in the proposed method, this paper first established a baseline model. This model only uses CCAM and DCRF to generate pseudo-labels for supervision and excludes the Multi-feature Aggregation Module (MFA) and the Context Feature Interaction Module (CFI), serving as the baseline model. Subsequently, this paper incrementally added the proposed modules to the baseline model and analyzed the contributions of each module in detail. As shown in the results in Table 2, each module introduced into the model plays a decisive role in achieving the final excellent performance. It can be concluded that the method proposed in this paper makes significant contributions to salient object detection.

Table 2. Ablation experiments on DUT-OMRON dataset

MOCO	DINO	PSU	MFA	CFI	$F_\beta \uparrow$	$E_m \uparrow$
✓	×	×	×	×	0.716	0.835
×	✓	×	×	×	0.663	0.793
✓	✓	×	×	×	0.726	0.835
✓	×	✓	×	×	0.727	0.838
✓	✓	✓	×	×	0.731	0.840
✓	✓	✓	✓	×	0.743	0.848
✓	✓	✓	✓	✓	**0.752**	**0.862**

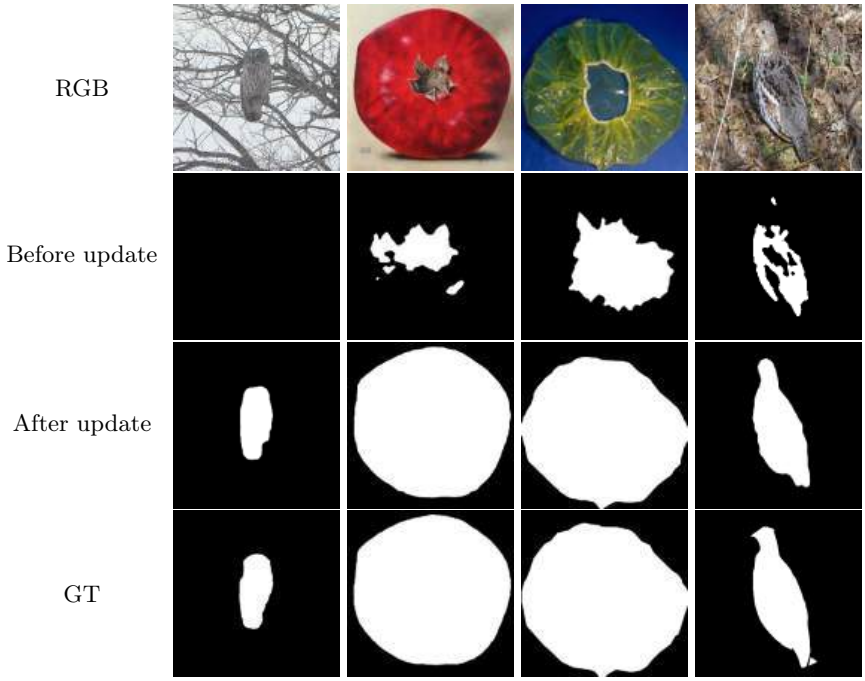


Fig. 6. Comparison of pseudo-labels before and after the update.

As shown in Figure 6, the visual differences between the pseudo-labels before and after updating are displayed. It is evident that the pseudo-labels updated using the self-updating method are closer to the ground-truth labels and better suited for the salient object detection task.

5 Conclusion

The comprehensive evaluation across multiple datasets demonstrates the robustness and effectiveness of the proposed method. Our approach consistently delivers competitive performance compared to both unsupervised and mainstream methods. Specifically, it matches the performance of fully-supervised and weakly-supervised methods on some datasets, while maintaining comparable results with mainstream methods on others. These findings highlight the potential of our method to bridge the gap between unsupervised and supervised learning in salient object detection. Future work will focus on optimizing the model architecture further and exploring its application in more diverse and complex scenarios.

References

1. Li, G., Xie, Y., Lin, L.: Weakly supervised salient object detection using image labels. In: AAAI, vol. 32, no. 1, pp. 7024–7031. AAAI Press, 2018.

2. Zhang, L., Zhang, J., Lin, Z., Lu, H., He, Y.: Capsal: Leveraging captioning to boost semantics for salient object detection. In: IEEE CVPR, pp. 6024–6033. IEEE, 2019.
3. Liu, Y., Wang, P., Cao, Y., Liang, Z., Lau, R.W.H.: Weakly-supervised salient object detection with saliency bounding boxes. IEEE TIP **30**, 4423–4435 (2021).
4. Zhang, J., Yu, X., Li, A., Song, P., Liu, B., Dai, Y.: Weakly-supervised salient object detection via scribble annotations. In: IEEE CVPR, pp. 12546–12555. IEEE, 2020.
5. Gao, S., Zhang, W., Wang, Y., Guo, Q., Zhang, C., He, Y., Zhang, W.: Weakly-supervised salient object detection using point supervision. In: AAAI, vol. 36, no. 1, pp. 670–678. AAAI Press, 2022.
6. Zhou, H., Chen, P., Yang, L., Xie, X., Lai, J.: Activation to saliency: Forming high-quality labels for unsupervised salient object detection. IEEE TCSVT **33**(2), 743–755 (2022).
7. Xie, J., Xiang, J., Chen, J., Hou, X., Zhao, X., Shen, L.: C2am: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. In: IEEE CVPR, pp. 989–998. IEEE, 2022.
8. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020).
9. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: IEEE International Conference on Computer Vision, pp. 9650–9660. IEEE, 2021.
10. Zhu, W., Liang, S., Wei, Y., Sun, J.: Saliency Optimization from Robust Background Detection. In: IEEE CVPR, pp. 2814–2821. IEEE, 2014.
11. Liu, J.-J., Hou, Q., Cheng, M.-M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. In: IEEE CVPR, pp. 3917–3926. IEEE, 2019.
12. Pang, Y., Zhao, X., Zhang, L., Lu, H.: Multi-scale interactive network for salient object detection. In: IEEE CVPR, pp. 9413–9422. IEEE, 2020.
13. Xu, B., Liang, H., Liang, R., Chen, P.: Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection. In: AAAI, vol. 35, no. 4, pp. 3004–3012. AAAI Press, 2021.
14. Liang, W., Ran, P., Bai, M., Liu, X., Githinji, P.B., Zhao, W., Qin, P.: External Prompt Features Enhanced Parameter-Efficient Fine-Tuning for Salient Object Detection. In: International Conference on Pattern Recognition, pp. 82–97. Springer, 2024.
15. Piao, Y., Wang, J., Zhang, M., Ma, Z., Lu, H.: To be Critical: Self-Calibrated Weakly Supervised Learning for Salient Object Detection. arXiv preprint arXiv:2109.01770 (2021).
16. Zhang, J., Yu, X., Li, A., Song, P., Liu, B., Dai, Y.: Weakly-supervised salient object detection via scribble annotations. In: IEEE CVPR, pp. 12546–12555. IEEE, 2020.
17. Piao, Y., Wang, J., Zhang, M., Lu, H.: Mfnet: Multi-filter directive network for weakly supervised salient object detection. In: IEEE International Conference on Computer Vision, pp. 4136–4145. IEEE, 2021.
18. Gao, S., Zhang, W., Wang, Y., Guo, Q., Zhang, C., He, Y., Zhang, W.: Weakly-Supervised Salient Object Detection Using Point Supervision. In: AAAI, 2022.
19. Nguyen, T., Dax, M., Mummadi, C.K., Ngo, N., Nguyen, T.H.P., Lou, Z., Brox, T.: Deepusps: Deep robust unsupervised saliency prediction via self-supervision. In: NIPS, vol. 32, 2019.

20. Yan, P., Wu, Z., Liu, M., Zeng, K., Lin, L., Li, G.: Unsupervised domain adaptive salient object detection through uncertainty-aware pseudo-label learning. In: AAAI, vol. 36, no. 3, pp. 3000–3008. AAAI Press, 2022.
21. Wang, Y., Zhang, W., Wang, L., Liu, T., Lu, H.: Multi-source uncertainty mining for deep unsupervised saliency detection. In: IEEE CVPR, pp. 11727–11736. IEEE, 2022.
22. Zhou, H., Qiao, B., Yang, L., Lai, J., Xie, X.: Texture-guided saliency distilling for unsupervised salient object detection. In: IEEE CVPR, pp. 7257–7267. IEEE, 2023.
23. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: IEEE CVPR, pp. 7479–7489. IEEE, 2019.
24. Pang, Y., Zhao, X., Zhang, L., Lu, H.: Multi-scale interactive network for salient object detection. In: IEEE CVPR, pp. 9413–9422. IEEE, 2020.
25. Yu, S., Zhang, B., Xiao, J., Lim, E.G.: Structure-consistent weakly supervised salient object detection with local saliency coherence. In: AAAI, vol. 35, no. 4, pp. 3234–3242. AAAI Press, 2021.
26. Fan, D.-P., Gong, C., Cao, Y., Ren, B., Cheng, M.-M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. arXiv preprint arXiv:1805.10421 (2018).
27. Chen, L.-C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017).
28. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: IEEE CVPR, pp. 1597–1604. IEEE, 2009.
29. Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: IEEE CVPR, pp. 733–740. IEEE, 2012.
30. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: IEEE CVPR, pp. 136–145. IEEE, 2017.
31. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: IEEE CVPR, pp. 1155–1162. IEEE, 2013.
32. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.-H.: Saliency detection via graph-based manifold ranking. In: IEEE CVPR, pp. 3166–3173. IEEE, 2013.
33. Li, G., Yu, Y.: Deep contrast learning for salient object detection. In: IEEE CVPR, pp. 478–487. IEEE, 2016.
34. Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P.: Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
35. Yang, T., Wang, Y., Zhang, L., Qi, J., Lu, H.: Depth-inspired label mining for unsupervised rgb-d salient object detection. In: ACM Multimedia, pp. 5669–5677. ACM, 2022.

Intelligent Compilation System for Chinese Character Animation Based on Dynamic Data Sets

Xin Luo[1,2,3][0009-0002-9559-1167] and Qingsheng Li[1,2][0000-0001-7617-9092]

[1] Communication University of Zhejiang, Hangzhou 310018, Zhejiang,China
[2] Hangzhou Liangyun Intelligent Technology Co.,Ltd, Hangzhou 310023, Zhejiang, China
[3] University College London,London WC1E6BT, UK
luoxin020228@163.com

Abstract. With the rapid evolution of AI and human-computer interaction technologies, Chinese character animation has gained wide applications in film/TV effects, digital education, and cultural heritage digitization. Current methods relying on static datasets face critical limitations in generation efficiency, style diversity, and real-time responsiveness. We address these challenges through an intelligent animation system incorporating three interconnected innovations: a dynamic dataset architecture supporting real-time updates for over 3,000 characters, a style-adaptive character description library, and a decoupled compilation-rendering framework that independently manages content generation and visual execution. By integrating stroke feature extraction with stroke-order reconstruction algorithms, our system automatically converts input characters into customizable animations with parametric control of curve smoothness and motion dynamics. Experimental validation confirms substantial efficiency improvements over conventional approaches, coupled with robust cross-platform compatibility and enhanced interactive capabilities across diverse usage scenarios. This work establishes a new paradigm for dynamic dataset-driven character animation systems.

Keywords: Chinese character animation generation, dataset, real-time rendering

1 Introduction

1.1 Background

With the rapid development of digital artificial intelligence and human-computer interaction technology, Chinese character animation, as a fusion form of visual communication and semantic expression, is showing important application value in many fields. In the scenes of film and television special effects production, digital education platform, digital human voice synchronous display, and cultural heritage visual display, the demand for fine and stylistically diversified Chinese character animation is becoming more and more urgent. For example, in film and TV post-production, anthropomorphic Chinese character animation can enhance the expressive

power of the screen; while in Chinese language education, dynamic Chinese character animation can help to improve learners' understanding of the stroke order, structure, and meaning, which is especially inspiring and interactive for non-native language learners.

However, most of the mainstream Chinese character animation generation methods rely on static datasets, such as existing stroke order character databases, standard vector data or fixed calligraphic style character databases. These datasets are mostly designed to serve the needs of standard Chinese character display at the early stage of design, and they have the following three core problems in animation generation:

1. Low generation efficiency: static data need to be parsed and interpolated to form animation sequences, which is difficult to meet the demand for real-time or batch generation.Recent studies have proposed methods like StrokeGAN to address efficiency issues by incorporating stroke encoding into generative models, thereby enhancing the generation process[1].

2. Single style: most datasets only support canonical writing styles and cannot express artistic and diverse expressions of Chinese characters [2].To overcome this limitation, approaches such as ZiGAN have been developed, enabling fine-grained Chinese calligraphy font generation through few-shot style transfer, thus allowing for a broader range of stylistic expressions [3].

3. Weak extensibility: it is difficult to extend new characters, variant characters or specific stroke styles, and the animation generation lacks universal adaptability.Innovative models like the one proposed by Chen et al. utilize generative adversarial networks to learn one-to-many stylized Chinese character transformations, enhancing the system's adaptability to new characters and styles [4].

To cope with the above problems, this paper focuses on the potential application of dynamic datasets in Chinese character animation. Dynamic datasets not only record stroke paths and time series information, but also integrate the stylistic features of different writers to achieve time-sensitive and expressive animation generation [5]. With the introduction of dynamic data in the generation system, it can be combined with deep learning models for real-time modelling and style migration, thus significantly improving the efficiency and diversity of animation generation[6].

For example, dynamic stroke libraries constructed based on online writing data have been widely used in handwriting recognition and personalised handwritten font synthesis, providing a theoretical and practical basis for the development of intelligent animation systems. There are also research attempts to apply Generative Adversarial Networks (GAN) or Transformer models to style migration and motion trajectory prediction of dynamic Chinese character stroke data, which have achieved preliminary results[2].

1.2 Background

This research aims to construct a dynamic dataset-driven intelligent compilation system for Chinese character animation for various application scenarios, such as film and television special effects, educational digitisation, and cultural communication. Specific objectives include: to propose a dynamic dataset architecture that supports real-time addition, deletion, and modification of Chinese characters, which can record

stroke paths, time series, and style information, and achieve efficient management and flexible expansion of Chinese character data; to design a set of intelligent compilation processes with dynamic data as the core input, which integrates structural parsing and animation generation models, to improve the system's generation efficiency and degree of intelligence, and have the capability of adapting to multi-terminal, multi-style, multi-context, and multi-directional animations. The system has the ability to adapt to multiple terminals, styles, and contexts.

The technical innovation of this research mainly reflects the dynamic incremental updating mechanism, which breaks through the limitation of static font and realises the real-time addition, deletion and style customization of Chinese character data to provide the data basis for the diversity animation.

2 Related Work

2.1 Chinese character animation generation technology

Currently, the generation of Chinese character animation can be mainly classified into three types of technical paths: keyframe-based interpolation, physical simulation methods, and deep learning generation.

Early animation production relied on keyframe interpolation, such as the use of commercial tools such as Adobe After Effects for manual keyframe annotation and path adjustment[7]. Although the accuracy of this method is high, the production cost is large, the generality is weak, and it is difficult to meet the needs of large-scale batch generation.

The physical simulation method simulates the writing process through physical engines such as Mass-Spring Model, which tries to restore the trajectory of the brush strokes on the physical level.

In recent years, deep learning techniques have been widely used in the field of Chinese character animation generation. Researchers have proposed a variety of models based on Generative Adversarial Networks (GAN) and Recurrent Neural Networks (RNN) for style migration and dynamic generation of Chinese characters. For example, the Auto-Encoder Guided GAN model proposed by Lyu et al. is able to convert standard fonts into calligraphic fonts with specific styles[8], which enhances the diversity and artistry of generated Chinese characters.

2.2 Dynamic data set techniques

Dynamic dataset management is one of the key technologies in the system supporting real-time generation of Chinese character animation. Traditional graphic data are mostly managed by static version, such as Git-LFS and other tools have basic tracking ability in image and font data management, but synchronisation delays and access conflicts often occur when dealing with large-scale and multi-version graphic files, which makes it difficult to meet the demands of real-time animation synthesis.

2.3 Existing challenges

Although some progress has been made in Chinese character animation generation and dynamic data management, there are still a number of technical bottlenecks in practical applications:

Firstly, there is a contradiction between data dynamics and animation stability. Frequent data updates may lead to unstable animation generation and style jumps, affecting visual coherence.

Secondly, the real-time compilation of large-scale glyph data is not efficient enough, and the existing methods are difficult to meet the immediate response requirements in multi-threaded concurrent environments, especially in educational platforms or interactive media that exhibit latency problems.

Finally, the consistency guarantee of cross-platform animation generation is still a difficult problem. In different terminals (e.g., Web, mobile, VR devices), the animation rendering mechanism varies greatly, resulting in the generation of results that are difficult to unify in terms of time synchronisation and visual style.

3 Approaches

3.1 Chinese character glyph description library

This study needs to extract the core data of Chinese characters based on the dynamic description library of Chinese character glyphs [9][10], and to organise and structure the data reasonably to ensure the efficiency and scalability of the subsequent processing.

The system adopts a structured glyph description library to store the stroke information of each Chinese character. Each Chinese character consists of multiple strokes, each stroke is represented by a series of two-dimensional coordinate points (x, y), and the storage format has been standardised to ensure the consistency of machine reading and parsing. The first 3 bits of the data file are the header information, describing the basic attributes of the character, and the 4th bit is the coordinate data of the strokes. Among them, the boundary point is marked by (-64, 0), which is used to separate neighbouring strokes, and the end point is marked by (-64, -64), which indicates the end of all the feature point data of the Chinese character. This glyph library supports accurate reconstruction at the stroke level and provides data support for subsequent dynamic rendering and interactive applications.

In the system, the glyph parsing process is realised by cyclically reading the data array, extracting and caching the coordinate points of each stroke, and drawing them as a continuous trajectory as soon as the stroke termination symbol is encountered. The advantage of this structure is that it can not only accurately restore the traditional writing process, but also provide the basic data unit for animation speed control.

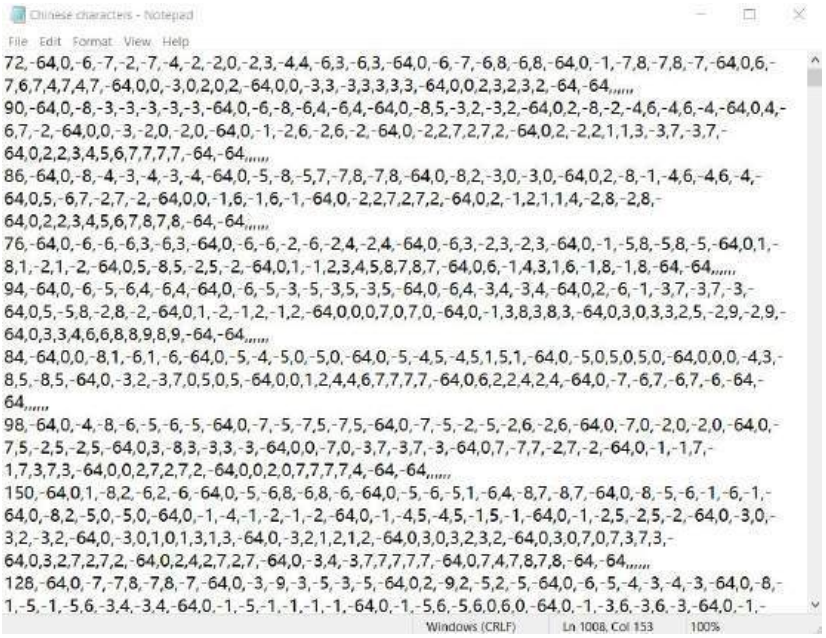


Fig. 1. Dynamic description library of Chinese character glyphs (partial)

3.2 Chinese character dynamic datasets

The animation generation module is based on the point-by-point drawing strategy, which realises the dynamic reconstruction and multi-speed playback of the stroke structure of Chinese characters. The system first receives the coordinate points of the characters and plots them in the MATLAB graphical interface. The plotting is done by the `plot(x, y)` command, which sets the line thickness and marker size to enhance the visual expression. The animation process is controlled by the `pause(t)` control interval, where the time can be set by the user to adjust the animation speed. The system supports switching from point-by-point drawing (slow demonstration) to instantaneous drawing of the whole stroke (fast presentation), which makes the animation suitable for both calligraphy and brushstroke teaching, as well as high-speed special effects generation and other application scenarios.

In the complete process, the system first receives the content of the Chinese character to be queried through the input interface, then loads its corresponding stroke data, decodes and draws it through the animation algorithm, and renders it stroke by stroke in the graphic window according to the set speed, and ultimately outputs a complete animation of the Chinese character writing process. The whole compilation and rendering process is completely automated, with good real-time response capability and user interaction experience.

4 Experiments

This system aims to build an experimental platform that can efficiently generate, dynamically invoke and intelligently render Chinese character animations. The overall process consists of three key steps: dataset generation, dynamic query and animation rendering. The following is a detailed description of the experimental process using the Chinese character "皑" as an experimental object:

Step 1: Generate dynamic Chinese character dataset

Firstly, the curvature of the basic character description model built into the system is adjusted by setting different curve control parameters, and dynamic data representations of more than 3,000 commonly used Chinese characters, including "皑" are generated in batch. Each Chinese character is indexed by a standard code, and each piece of data consists of feature points arranged in stroke order and automatically labelled with the corresponding Chinese character for subsequent retrieval.

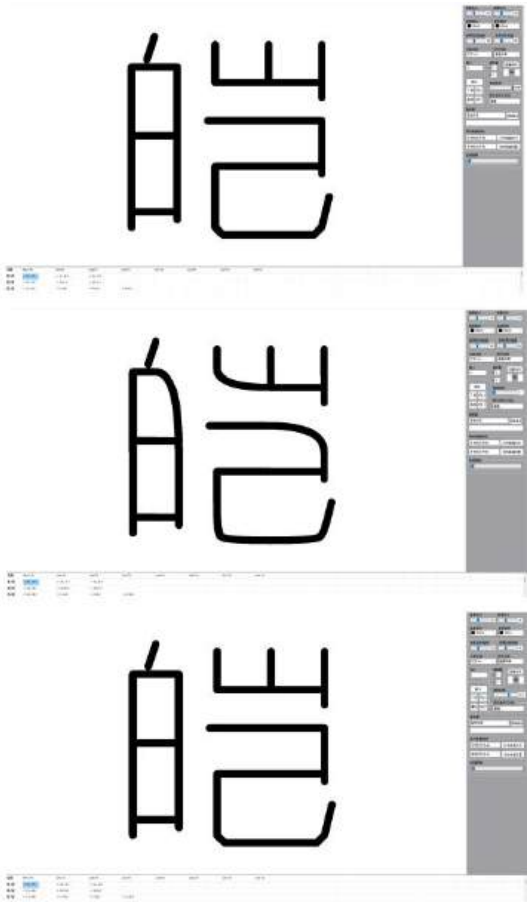


Fig. 2. Dynamic Chinese character datasets generated by different curves ("皑")

Assuming that the feature point data of Chinese character "皑" in the generated dynamic dataset is variable C, which represents the feature point information of Chinese character "皑":

C =

"104,-64,0,-10,-14,-11,-11,-11,-11,-64,0,-13,-10,-13,11,-13,11,-64,0,-13,-10,-7,-10,-7,10,-7,10,-64,0,-13,-1,-1,-1,-1,-1,-1,-1,-7,-1,-64,0,-13 9,-7,9,-7,9,-64,0,5,-13,5,-8,5,-8,-64,0,-2,-13,-2,-8,12,-8,12,-8,-64,0,12,-13,12,-6,12,-6,-64,0,-3,-3,12,-3,12,-3,12,4,12,4,-64,0,-2,3. 12,3,12,3,-64,0,-2,3,-2,11,-1,12,11,12,12,11,13,7,13,7,-64,-64 ,,,,,"

The system automatically parses the data, separates the strokes, and writes them into the dynamic dataset in a standardised format, realising a basic data system for Chinese character animation with a clear structure, rapid retrieval and flexible updating.

Step 2: Dynamic reading of Chinese character data

When the user inputs the Chinese character " 皑 " into the system, the system immediately locates the corresponding entry in the glyph description library through the tag search mechanism and reads the complete coordinate point data of the Chinese character "皑".

```
Please enter the Chinese character you want to query: 皑
Chinese character: 皑
complete description of the data:
,"104,-64,0,-10,-14,-11,-11,-11,-11,-64,0,-13,-10,-13,11,-
```

Fig. 3. Dynamic reading of the data of the Chinese character "皑"

Step 3: Separation of Chinese Strokes

The reading module parses the stroke structure according to predefined rules, where:

The feature point marker (-64, 0) indicates the end of a stroke;

The feature point (-64, -64) indicates the termination of the entire word;

All point information is organised in stroke order and temporarily stored in set S.

For example, the parsed set S is shown below (only partially):

S = {

s0 = {(-10,-14), (-11,-11), (-11,-11)},

s1 = {(-13,-10), (-13,11)},

s2 = {(-13,-10), (-7,-10), (-7,10)},

...

}

This data provides the raw graphic path data for subsequent animation drawing modules.

Step 4: Generate Chinese Character Animation

The animation generation module receives the stroke data in the set S and generates animation by drawing the strokes one by one in accordance with the stroke order. The system performs the following operations for each stroke:

1. Extract all the coordinate points of the current stroke from the set S;
2. Based on the number and arrangement of points:
   - If the number of points is two, connect them directly with a straight line;
   - If the number of points is more than three, the path is fitted using a polyline or smooth curve;
3. Draw each stroke frame by frame at a set speed to form an animation.

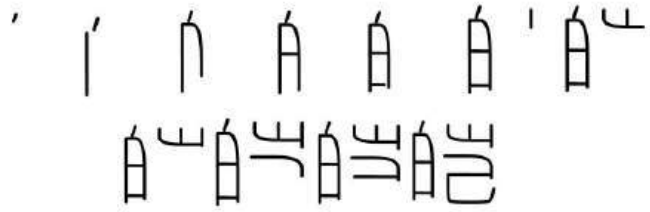


Fig. 4. Animation of Chinese character strokes ("皑")

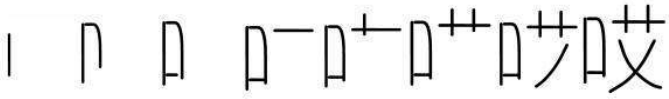


Fig. 5. Animation of Chinese character strokes ("哎")

Plotting is implemented using the MATLAB graphics engine or equivalent graphics APIs, and rendered in real time using the `plot` function; the tempo of each stroke can be controlled by setting the speed of the plot, thus enabling flexible switching between slow teaching and fast visual presentation.

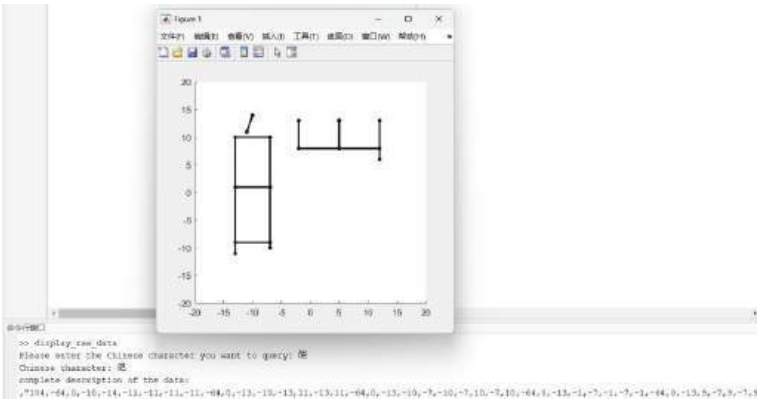


Fig. 6. Searching Chinese character "皑" and generating Chinese character animation in real time.

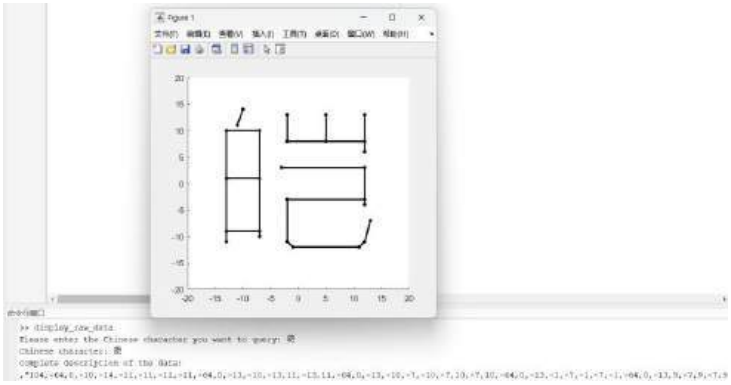


Fig. 7. End of animation generation of Chinese character "皑"

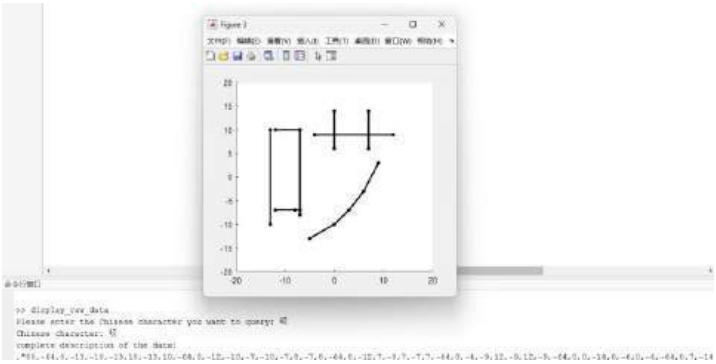


Fig. 8. Searching Chinese character "哎" and generating Chinese character animation in real time.

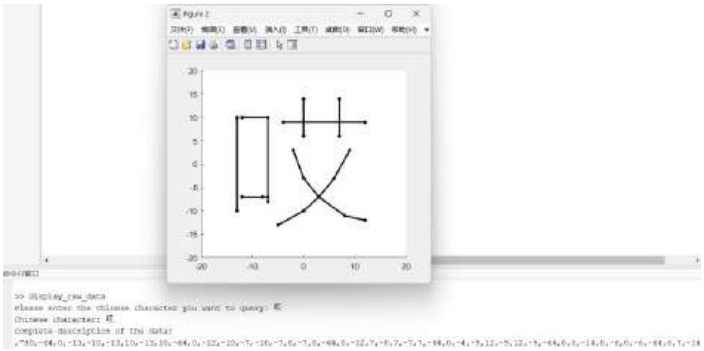


Fig. 9. End of animation generation of Chinese character "哎"

In the end, the system synthesises all the strokes in S into the dynamic structure C of a complete Chinese character, and realises the animated writing display of the Chinese character "皑". This process supports efficient rendering, speed adjustment,

and real-time user interaction, and provides a technical foundation for subsequent deployment in multiple scenarios such as educational platforms and special effects engines.

In order to enhance the robustness and versatility of the system, the system constructs specialized stroke data sets for complex font styles such as cursive and semi-cursive.

Cursive script dataset: it is collected from the cursive script glyph database, covering a large number of characteristic strokes such as continuous strokes, omissions, deformations and so on. The system extracts the connection patterns between strokes through the trajectory deconstruction algorithm, and introduces the curvature analysis and stroke direction modeling technology to restore the free-flowing and rhythmic dynamic strokes of cursive script;

Semi-cursive Script Data Set: Constructing a transitional font style based on cursive writing, taking into account the actual writing characteristics of structural normality and stroke deformation. The system can combine the standard stroke template with the actual writing trajectory, automatically determine the starting and stopping point fuzzy situation, and carry out flexible stroke synthesis and dynamic fitting to ensure the balance between style restoration and recognition.

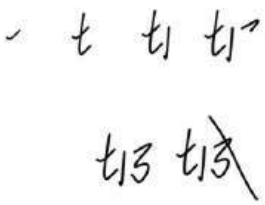


Fig. 10. Animation of semi-cursive strokes("城")

Fig. 11. Animation of semi-cursive strokes("橙")

Fig. 12. Animation of cursive strokes("啊")

5 Conclusion

This study demonstrates the clear advantages of an intelligent Chinese character animation system based on a dynamic dataset in terms of generation efficiency, scene adaptability, and user interaction. Leveraging a decoupled compile-render architecture and dynamic data-driven approach, the system achieves high efficiency, with an average generation time of just 0.15 seconds per character. It also supports flexible control of curve parameters and animation speed, enabling smooth adaptation to various scenarios—such as digital education, film effects, and human-computer interaction—while maintaining consistent performance across platforms like web and mobile.

Acknowledgments. This study was funded by Key R&D Project of Zhejiang Province,China (grant number 2021C03137).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Zeng, J., Chen, Q., Liu, Y., Wang, M., & Yao, Y. (2020). StrokeGAN: Reducing Mode Collapse in Chinese Font Generation via Stroke Encoding. AAAI Conference on Artificial Intelligence.
2. Pan Zeyu. A method for generating Chinese calligraphy fonts based on generative adversarial networks [D]. Hangzhou Dianzi University, 2023. DOI: 10.27075/d.cnki.ghzdc.2023.000953.
3. Wen, Q., Li, S., Han, B., & Yuan, Y. (2021). ZiGAN: Fine-grained Chinese Calligraphy Font Generation via a Few-shot Style Transfer Approach. Proceedings of the 29th ACM International Conference on Multimedia.
4. Chen, J., Ji, Y., Chen, H., & Xu, X. (2019). Learning one-to-many stylised Chinese character transformation and generation by generative adversarial networks. IET Image Process., 13, 2680-2686.
5. Li, Y. , Yang, Q. , Chen, Q. , Hu, B. , Wang, X. , & Ding, Y. , et al. (2023). Fast and robust online handwritten chinese character recognition with deep spatial and contextual information fusion network. Multimedia, IEEE Trans. on (T-MM), 25(000), 13.
6. WangZi-Rui, DuJun, & WangJia-Ming. (2020). Writer-aware CNN for parsimonious HMM-based offline handwritten Chinese text recognition.
7. Burtnyk, N., & Wein, M. (1998). Interactive skeleton techniques for enhancing motion dynamics in key frame animation. Seminal graphics: pioneering efforts that shaped the field.
8. Lyu, P. , Bai, X. , Yao, C. , Zhu, Z. , Huang, T. , & Liu, W. . (2017). Auto-Encoder Guided GAN for Chinese Calligraphy Synthesis. arXiv.
9. Xiong, J., Liu, X., & Li, Q. (2015). Ontology Description of Chinese Character Semantics. 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, 709-713.
10. Qingsheng, Li., Quan, Liu.(2015). A Novel Dynamic Description and Generation Method for Chinese Character. 719-724. doi: 10.1109/CIT/IUCC/DASC/PICOM.2015.1