

Received 31 January 2023, accepted 17 February 2023, date of publication 22 February 2023, date of current version 27 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3247627

## RESEARCH ARTICLE

# A Vision and Semantics-Jointly Driven Hybrid Intelligence Method for Automatic Pairwise Language Conversion

JIANCHAO ZHAO<sup>1</sup>, ZHIZHI FENG<sup>2</sup>, AND ZHENG DU<sup>3</sup>

<sup>1</sup>Henan Polytechnic Institute, Nanyang 473000, China

<sup>2</sup>School of Humanities and Arts, Chongqing University of Science and Technology, Chongqing 401331, China

<sup>3</sup>Baidu Online Network Technology Company, Beijing 100000, China

Corresponding author: Jianchao Zhao (zjcl@sina.com)

This work was supported by the Henan Polytechnic Institute.


**ABSTRACT** In modern society where connections among nations have been more and more frequent, it remains important to realize automatic language conversion methods for the public. Currently, most the existing research works were conducted upon the basis of semantics analysis. But from the perspective of linguistics, the vision characteristics is also a kind of concomitant existence. To deal with such challenge, this paper proposes a vision and semantics-jointly driven hybrid intelligence method for automatic pairwise language conversion. The whole technical framework can be divided into two components: vision sensing part and semantics sensing part. For the former, the virtual reality is introduced for use to capture the visual feature representation for language contents. For the latter, the recurrent neural network model is utilized to capture semantic feature representation for language texts. They are then integrated into a jointly driving framework, so as to improve the conversion efficiency. Taking two dialects (Sichuan dialect and Chongqing dialect) in China as the example, the simulative experiments are conducted on massive real-world training corpus to evaluate the proposal. The results can reflect feasibility of it.

**INDEX TERMS** Hybrid intelligence, virtual reality, deep neural network, automatic language conversion.

## I. INTRODUCTION

In the age of digital devices, we can achieve better learning through technology [1]. Virtual reality seems to be the next step in the development of education, where learners can learn the language whenever and wherever they want to improve their language skills, which will be the inevitable trend in the future of education [2]. Information technology is a tool for education, and Chacon suggests that virtual reality will be a powerful tool for the classroom based on the Tower of Experience theory [3]. Students can use immersive virtual reality to complete tasks twice as fast as students using traditional computer programs [4]. This indicates that more and more scholars are focusing on the efficient linkage between VR technology and education [5]. In addition, a national

survey found that 90% of educators believe virtual technology is an effective way to provide differentiated and personalized learning experiences for students [6]. Virtual reality technology has matured with the continuous innovation of modern computer technology and the constant update of 3D display devices [7]. The immersive Ness, interactivity, and flexibility of virtual reality make it possible to build virtual scenarios for teacher training, providing practitioners with highly simulated training scenarios to help them train their language skills [8], [9]. Based on the fully immersive virtual reality device, practitioners can prepare for different needs and training difficulties, breaking through the traditional time and space constraints and the disparity in teacher-student ratio so that students can arrange their own time to practice freely and have more training opportunities [10]. In organizing the literature, we found that many scholars study the specific implementation of virtual reality scenes in the process [11].

The associate editor coordinating the review of this manuscript and approving it for publication was Laura Celentano .

Still, there is little research on the application effect of virtual reality scenes [12]. Still, education, not technology alone, should be how to use technology reasonably and effectively to systematically and scientifically solve the practical problems faced in education, to be learner-centered, from the learner, to help learners how to properly use information technology to assist their learning, and to use information technology to promote the development of the learners themselves [13].

With the continuous development of today's society, the regional cultures of various provinces are gradually emerging as a trend of intermingling [14]. The promotion and application of Mandarin have facilitated cultural exchange between regions but have also led to the loss of dialect culture. At the same time, international respect for cultural diversity has been proposed, and the preservation of dialects has become a significant trend of the times [15]. Today, dialects are finding a new direction in the design field through visual design. Regarding dialects, "accent" is the first thing we can think of, and the highly recognizable vernacular is dialects' most direct external expression [16]. By converting dialects from "accents" to visual images and using visual language to close the distance between dialects, we can not only guide people to understand a city's dialect culture and increase the fun and readability of dialects but also use visuals local customs, charming scenery, and exceptional food, promoting the exchange and dissemination of different city cultures, communication, and dissemination between cultures [17].

In today's rapidly developing society, people should pay attention to new things and weather and inherit and protect the culture of regional dialects inscribed in their bones. As a kind of intangible cultural heritage, we can try to interpret the regional dialects in a "visible" way and use some specific graphic symbols to start the communication of information to prevent the transmission barriers and phonetic and semantic barriers caused by the regional dialects in the process of transmission, and to reduce the communication barriers when people from different regions speak the dialects. The purpose is to prevent the communication and phonological and semantic barriers caused by regional dialects in the transmission process and to reduce the communication barriers when people speak dialects in different regions. This paper uses the language of graphic symbols to translate regional dialects, combining sound, meaning, and form to give people a novel visual perception. The use of graphics to express the dialect enhances the flavor of life. It enriches the content of the idiom, transforming it from "hearing" to "seeing," conveying information and improving interest. The graphic design explains the local dialects, allowing people from other places to break through the language barrier of the local culture and deepen their understanding of the local language and culture.

The main content of the article is how to use our professional knowledge to better translate the "language symbol" dialect into "visual symbol" graphics, which is the main content of the subject to be further researched. The goal of the research is to better display the visual language in the

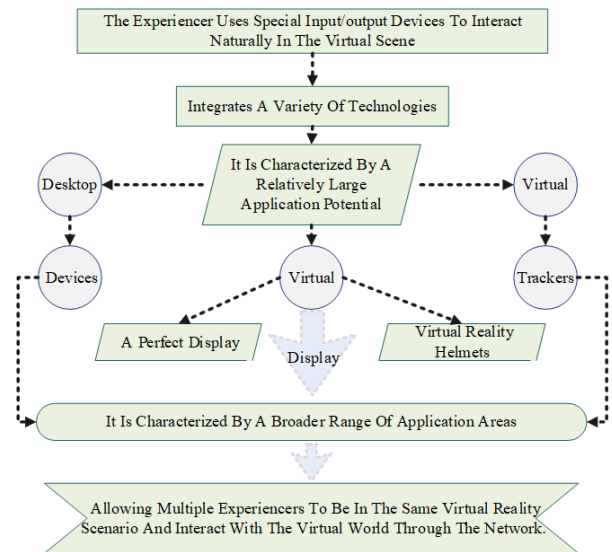


FIGURE 1. Virtual reality technology scene recognition process.

design works, so that the dialects of Sichuan and Chongqing regions, with the support of regional culture, can increase people's sense of regional identity, make people have a sense of belonging to the city of Sichuan and Chongqing, and promote the spread of dialect culture, so that more people can understand the dialects of Sichuan and Chongqing and realize the importance of preserving regional culture.

## II. RELATED WORKS

Research on virtual reality has been going on for many years; it can be traced back to the 1940s, when it was applied to the military and the simulation training of astronauts. Currently, the application of virtual reality is becoming more and more widespread, from the beginning only in military and aerospace applications to now able to be applied to scientific research, such as experiments with aircraft in simulated flowing air, and industrial processing and manufacturing, such as the use of virtual reality to affect the process of machining parts [18]. It can also be applied in education, such as simulating the operation of volcanic eruptions so that students can visualize a new understanding of volcanic eruptions. From the current point of view, real-time and dynamic is the focus of virtual reality technology. Most of the research on virtual reality technology is based on Second Life as an example to analyze the teaching of the Chinese language; the VRML language is rarely used in development research [19].

Virtual environments provide the necessary elements with factual context, and language learning occurs by simulating contextual contexts. Many researchers are using virtual worlds as a platform for language learning; the Language Project Research Office is developing VR language learning projects [20]. Now, language teaching research is focused on projects on the Second Life platform, with a focus on the process of designing tasks and the effects of virtual reality

in the language learning process. In terms of self-efficacy in language skills, many overseas researchers have concluded that virtual Chinese language courses can help improve self-efficacy in language skills [21]. A two-year study found that the virtual course's Chinese lessons helped enhance students' self-efficacy and significantly helped learners who lacked real-life language background to master real-life experiences in the language.

The study Second Life (SL) could effectively improve spoken Chinese output in terms of language output research [22]. The purpose of the study was to determine the effectiveness of Second Life (SL) in enhancing overseas learners' spoken Chinese output. The usefulness of Second Life in improving spoken language output was evaluated through two phases of the study. One of the most successful aspects of virtual technology in designing Chinese language courses in a virtual environment is its application in education and training [23]. Three universities are collaborating on an innovative 3D visual design project in one of the most popular virtual worlds, Second Life. The project requires Second Life software for co-creation and collaborative design, and virtual technology provides powerful technical support for visualization education. As seen from the above research, many scholars or technologists have focused on creating virtual Chinese learning environments, exploring the impact of virtual environments on improving teaching and learning.

In the area of graphic design, Huda R researches the visualization design of the Henan dialect, not only discussing the need to bring more diversified and exciting interactive experiences to the communication and development of dialects in the new media environment but also studying the artistic forms of color and composition of the wood-paneled New Year paintings in Zhuxian Town and applying them to the micro-motion graphic design of Henan dialect, which provides new ideas on the way to combine traditional art and dialect visualization [24]. Hatipoglu B and Yilmaz C M explain the advantages of merging graphic design with dialect culture, discuss the necessity of graphic design for northeastern dialects, and list the design methods and introductory presentation forms of visualized graphics for dialects, which opens new directions for the combination of dialects and graphic design [25]. Cong R summarized the cases of dialect graphic design from three perspectives of context, semantics, and symbols and outlined three forms of transformation of sound and structure, meaning and form, and form and form to present dialect graphic design, expanding the ideas and methods of dialect graphic design [26].

In terms of typeface design, Loubeyre P and Occelli F from the College of Humanities and Arts of Hunan International Economics College take Changsha local dialect as the representative, use typeface design as the content expression, and aim to improve the additional cultural value of cultural and creative products and explore the method of expressing Changsha dialect on artistic and innovative products through typeface design, which is an essential reference value for the study of the combination of regional language, typeface

design, and cultural and creative products [27]. The research on dialect visualization is based on regions with outstanding regional dialect characteristics, while the research on the visualization of Sichuan and Chongqing dialects is still in a blank stage [28]. At present, the research on dialect visualization is based on regions with outstanding regional dialect characteristics, while the research on the visualization of Sichuan and Chongqing dialects is still in a blank stage. By combining graphic design and typeface design in visual communication, the study will present the dialects in the form of visuals on the innovative products, the cultural value of Sichuan and Chongqing dialects, eliminating the language barrier of dialect communication, and providing new ideas for the transmission and development of Sichuan and Chongqing dialects.

### III. THE PROPOSED METHOD

The research method of this paper uses the collection of visual graphic design works of different regional dialects for classification and comparison, and uses the theoretical knowledge of semiotics, design and linguistics to summarize and conclude the strengths and weaknesses of the current visual graphic design works of regional dialects, and selectively analyze and review them in this paper, and propose feasible design ideas for visual translation design of dialects in Sichuan and Chongqing.

#### A. VIRTUAL REALITY TECHNOLOGY MODEL DESIGN

Virtual Reality (VR) technology, as a category of computer simulation technology, is a computer technology as the core to be able to create and experience virtual environments. The experienter uses special input/output devices to interact naturally in the virtual scene. Virtual reality technology is a comprehensive technology that integrates a variety of technologies, including computer graphics technology and human-computer interaction technology and sensor technology, etc. Using these technologies, virtual reality technology can simulate not only the construction of scenes in natural environments but also build scenes that are difficult to experience in natural environments [29]. According to the different display methods, virtual reality technology can be divided into four types. Namely, virtual desktop reality, distributed virtual reality, augmented virtual reality, and immersive virtual reality.

(1) Desktop virtual reality: It is characterized by a broader range of application areas and is relatively easy to implement. Ordinary flat display devices are used to view virtual scenes, and users interact with virtual world objects through input devices such as mouse and keyboard. The relevant devices are computers, primary graphics workstations, projectors, keyboards, mice, etc. Although users can see three-dimensional objects, desktop virtual reality systems cannot immerse the experienter in them. (2) Distributed virtual reality: It is characterized by a wide range of application prospects and is suitable for multi-person interaction scenarios. It is a combination of computer network technology and

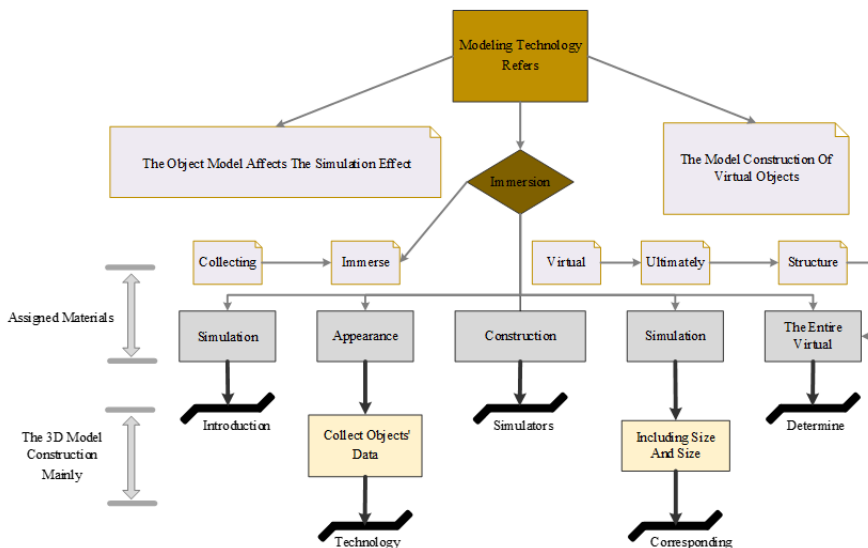


FIGURE 2. Virtual reality interaction design process.

virtual reality technology, allowing multiple experiencers to be in the same virtual reality scenario and interact with the virtual world through the network. The equipment used is the display, processing system, communication equipment, etc. (3) Augmented virtual reality: It is characterized by a relatively large application potential and convenient experience equipment. It is the virtual reality simulation of virtual objects superimposed on the natural world so that the real and virtual worlds become one. Therefore, it generally does not have high requirements for hardware configuration. Augmented virtual reality generally relies on mobile terminal devices. Augmented virtual reality technology often uses virtual reality to describe the evolution of the natural world and enhance people’s perception of real-world objects. (4) Immersive virtual reality: Its characteristics have a perfect display of virtual reality effects, able to multiple sensory information simultaneously to act on the user so that the user in the role of multi-sensory information stimulation to produce the same effect as the real world. The equipment used is virtual reality helmets, data gloves, and position trackers. The user enters the virtual reality world with the help of virtual reality devices and experiences the feeling of being in the virtual world. In contrast, the user uses sensor devices such as data gloves and position trackers. The process of virtual reality technology scene recognition is shown in Figure 1.

The leading virtual reality technologies include three parts of the content: modeling technology, display technology, and interaction technology.

(1) Modeling technology: modeling technology refers to the process of three-dimensional restoration of objects. The model construction of virtual objects is essential to virtual reality technology. There is no way to build virtual scenes without the support of three-dimensional models. The degree of refinement of the object model affects the simulation effect

of the whole virtual set. The model building process is to collect objects’ data first, including size and size, color, picture material, etc. Then the basic model is built, the basic model is assigned materials, mapping, and finally, the 3D model is rendered and exported. The success of the 3D model construction mainly depends on whether the simulation is high and whether the appearance is beautiful, and the structure of the 3D model of the object is ultimately related to whether the entire virtual scene can immerse the user. This requires collecting detailed information about the model before creating the object model, such as the size of the actual object, pictures, colors, etc.

(2) Display technology: Stereo display is one of the critical technologies of virtual reality, which gives people a stronger sense of immersion in the virtual world, and the introduction of the stereo display can make the simulation of various simulators more realistic. Therefore, it is necessary to study the stereo imaging technology and use the existing computer platform, combined with the corresponding software and hardware systems, to display the stereo view on a flat display. At present, stereo display technology mainly uses stereo glasses and other auxiliary tools to view stereo images.

(3) Interaction technology: The interaction technology of the virtual scene is mainly through collision detection, input device control, and location jumping to realize the interaction between the user and the three-dimensional scene, and the experiencer can use the HTC VIVE handle controller to both select and determine the location and turn the pages of the exercise material. Students will jump to the inside of the scene after selecting the setting and regret jumping to the selected scene interface by clicking the back button in the scene through the joystick controller. The virtual reality interaction design flow is shown in Figure 2.

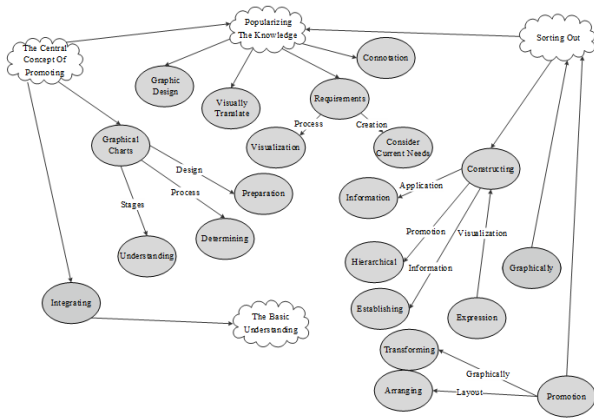


FIGURE 3. Visual transformation process of Sichuan and Chongqing dialects.

**B. CONSTRUCTION OF A VISUAL TRANSFORMATION MODEL FOR SICHUAN AND CHONGQING DIALECTS**

The graphic design of Sichuan and Chongqing culture information is based on the central concept of promoting and popularizing the knowledge of Sichuan and Chongqing culture, sorting out, constructing, and mining Sichuan and Chongqing culture information found on the basic understanding of Sichuan and Chongqing culture, and using the expression of graphic design to visually translate the complex and challenging cultural information into easy-to-understand graphical charts to improve the efficiency of readers’ information understanding. By understanding Sichuan and Chongqing culture, we create visualization works that meet the requirements of Sichuan and Chongqing aesthetic connotation and consider current needs [30]. Before the design creation begins, the design process needs to be clarified, as shown in Figure 3, which encompasses three experimental stages. First, the design preparation stage: determining the theme, understanding, and integrating the cultural information, and constructing the hierarchical order of artistic information; second, the design creation stage: establishing the visual expression style, graphically transforming the unit information, and arranging the shape; finally, the design promotion stage, graphically transforming the Sichuan and Chongqing cultural information for derivative application promotion design.

For the chapter filled with a large amount of information and data content to enter the unit information visualization conversion, we need to use the board layout, graphics and diagrams, and other means; the creation should fully comply with the principle of cultural and historical materials as the basis, the pursuit of information clarity, accuracy, and depth, for the construction of the overall bias visualization solid foundation. With the rapid development of the Internet, emoji packs have become an essential product of online culture. The emergence of emoji packs is undoubtedly an excellent way to fill this gap, greatly enhancing the emotional element in the process of modern Internet interaction and playing a role in

bringing friends and relatives closer. Thus, the combination of emoji packs and dialect has become a hilarious form of entertainment in the era of network communication. This kind of expression combining graphics and words is excellent for avoiding misunderstandings between the two parties in the communication process and creating a beautiful, harmonious, and witty interactive environment.

A backpropagation algorithm through time used by neural networks consists of two stages: forward propagation and backward propagation. First, let’s analyze the computation of forward propagation, assuming that there is an input sequence  $\{x^{<1>}, x^{<2>}, x^{<3>}, \dots, x^{<Tx>}\}$ , then use  $x^{<1>}$  and  $a^{<0>}$  to calculate the activation term for time step 1, then use  $x^{<2>}$  and  $a^{<1>}$  to calculate  $a^{<2>}$ , then calculate  $a^{<3>}$ , etc., until  $a^{<Tx>}$ . After calculating  $a^{<1>}$ , some parameters are needed  $W_a$  and  $b_a$ , which are used at each subsequent time step to calculate  $a^{<2>}$ ,  $a^{<3>}$ , etc. All these activation terms depend on the parameters  $W_a$  and  $b_a$ . With  $a^{<1>}$ , the neural network can calculate the first prediction  $y^{<1>}$ ; then, at the next time step, it continues to compute  $y^{<2>}$ ,  $y^{<3>}$ , etc., all the way to  $y^{<Ty>}$ .  $y$  is computed with parameters  $W_y$  and  $b_y$ , which will be used for all these nodes. A loss function is also needed to calculate the backpropagation, starting with the definition of an essential loss function.

$$L_{(y^{<t>})} = \sum \log y^{<t>} - \frac{\sqrt{(1 + y^{<t>})}}{\sqrt{\log(1 - y^{<t>})}} \tag{1}$$

where  $y$  corresponds to a specific word in the sequence, and if it is the name of someone, then the value of  $y^{<t>}$  is 1. The neural network will then output the probability value that the word is the name, say 0.1. Define it as the standard logistic regression loss function, also called Cross-Entropy Loss, which is the value of the loss function on a single position or a certain This is the loss function of the predicted value of a word at a single location or at a time step  $t$ . Now to define the loss function for the whole sequence, define  $L$  as:

$$L_y = \sum_{t=1} T_x - l^{<t>} + \frac{y^{<t>}}{y^{<t>} - 1} \tag{2}$$

In this calculation, the corresponding loss functions can be calculated by  $y^{<1>}$ , so the loss function of the first-time step, the loss function of the second time step, and so on are calculated until the last time step. Finally, to calculate the overall loss function, they are all added up, i.e., the loss function of each time step is added up. GRU is a model that can also handle sequential data and is a type of recurrent neural network. Still, it is also a variant of LSTM, and the reason why we learned it is because we found that many areas can be streamlined and improved after understanding LSTM, such as its complex model structure, so GRU was born. The internal structure of LSTM has been simplified, and the accuracy has been improved. Unlike the LSTM, the GRU does not introduce additional memory units, only updates gates to control how much information the current state needs to retain from historical messages and how much new information to receive from candidate states. The GRU achieves comparable

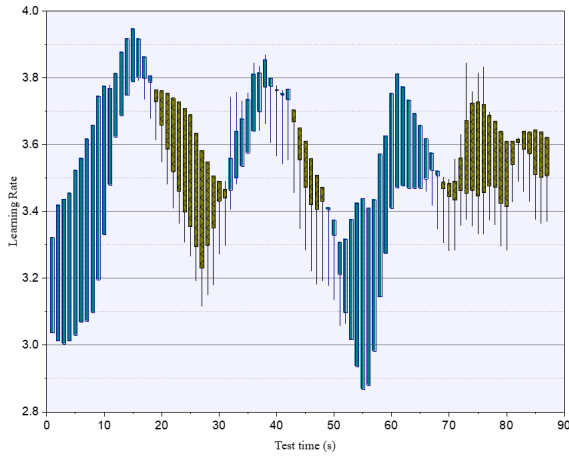


FIGURE 4. Parameter testing of network training.

results, improves training efficiency, and speeds up training time.

The input-output structure of the GRU is the same as that of a normal RNN; combining the current moment input  $x_t$  and the hidden state  $h_{t-1}$  passed down from the last moment, a candidate hidden state  $h_t$  is obtained by resetting the gate, and  $h_t$  mainly contains the information of the current input  $x_t$ . And the real hidden state  $h_t$ , which needs to be computed, will be obtained by passing down  $h_{t-1}$ , and the candidate hidden state  $h_t$  from the last moment, in which  $r_t$  is the gate that controls the reset and  $z_t$  is the gate that controls the update. The reset gate is used to manage the degree of forgetting previous information, which mainly determines how much past information needs to be overlooked; the smaller the value, the more information is omitted, and the reset gate can make the hidden state forget any information that is found irrelevant to the prediction in the future. The update gate is used to control how much information from the previous moment is passed into the current state; it determines how much past information is to be passed into the future; the more significant its value, the more information is passed in, the update gate control of the information in the previous hidden state can remember the long-term data and can reduce the risk of gradient disappearance. The calculation formula is as follows:

$$R_t = \int \frac{u_r - h_t + 1}{\sigma - w_r + x} \quad (3)$$

$$Z_t = \int \frac{(\sigma w_z + x) \times u_z h_{t-1}}{w_z x} \quad (4)$$

$\sigma$  is the sigmoid activation function, and the final value ranges from 0 to 1.  $w$  and  $u$  are the weight matrices to be learned,  $x$  is the input, and  $h_{t-1}$  is the hidden state at the last moment.  $h_t$  is the candidate's hidden state, which can be considered as new information at the current moment and is calculated as follows:

$$h_{t-1} = \int \tanh \frac{u_r \times h_{t-1}}{\sqrt{w_x - u_r}} \quad (5)$$

The candidate's hidden state is only related to the input, the hidden state of the previous moment.  $h_t - 1$  is related to the reset gate  $r - t$ , and  $r_t$  takes values from 0 to 1. If it tends to 0, then the information of the previous moment is forgotten. The hidden state is calculated as follows:

$$h_t = \sum_{t=1} \left( \frac{h_t}{z_t + z_t h_t} - \frac{h_{t-1}}{z_t + h_{t-1}} \right) \quad (6)$$

From Equation (6), the current hidden state depends on the hidden state of the previous moment and the candidate's secret state. If  $z_t$  tends to 0, it means that the information of the last moment is forgotten, and if  $z_t$  tends to 1, it means that the input information of the current moment is forgotten.

Creative design is a unique visual language and symbols different from other forms of expression, which can be repeatedly viewed through various channels and different audience groups and is the most convenient and intuitive presentation effect for disseminating information. The use of creative design as an artistic technique can make people's daily life more beautiful, and the level of aesthetic appreciation is relatively greatly improved; to create an excellent creative design, works must have a rich cultural connotation and eye-catching picture effect. Visual creativity is one of the core elements of innovative design, which can and effectively convey the information that the designer wants to express; rich in uniqueness and creativity, this fresh and exciting expression and form of expression tells people the inner world of the designer in a profound and visually impactful way of presentation, attracts attention and interest with a novel and original story image, so that people can firmly remember [31]. It will make people remember and think by looking up textual materials and design works related to Sichuan and Chongqing dialects, Sichuan and Chongqing culture and graphic design of dialects, and exploration of visual conversion of dialects, covering various aspects of visual communication, local dialects, folk culture, graphic symbols, etc., I learned and realized anew that behind the familiar and seemingly ordinary idioms, many cultural backgrounds and development histories are waiting for us to learn and explore. After compiling and understanding the knowledge of Sichuan and Chongqing dialects, I chose local people of different age groups who know the most about Sichuan and Chongqing dialects, as well as teachers and students around me and conducted research and exchanges with them to understand the local dialect culture of Sichuan and Chongqing from the perspective of local people, and to collect corpus and their opinions and suggestions on combining the vocabulary of Sichuan and Chongqing dialects with modern creative design.

#### IV. RESULTS

##### A. ANALYSIS OF THE DIALECT VISUAL TRANSFORMATION METHOD OF VIRTUAL REALITY TECHNOLOGY

Based on the GRU gated cyclic unit and Transformer language model, the model architecture for Sichuan and Chongqing dialect speech recognition is built. In general, the model adopts the basic GRU computing module, first extracts

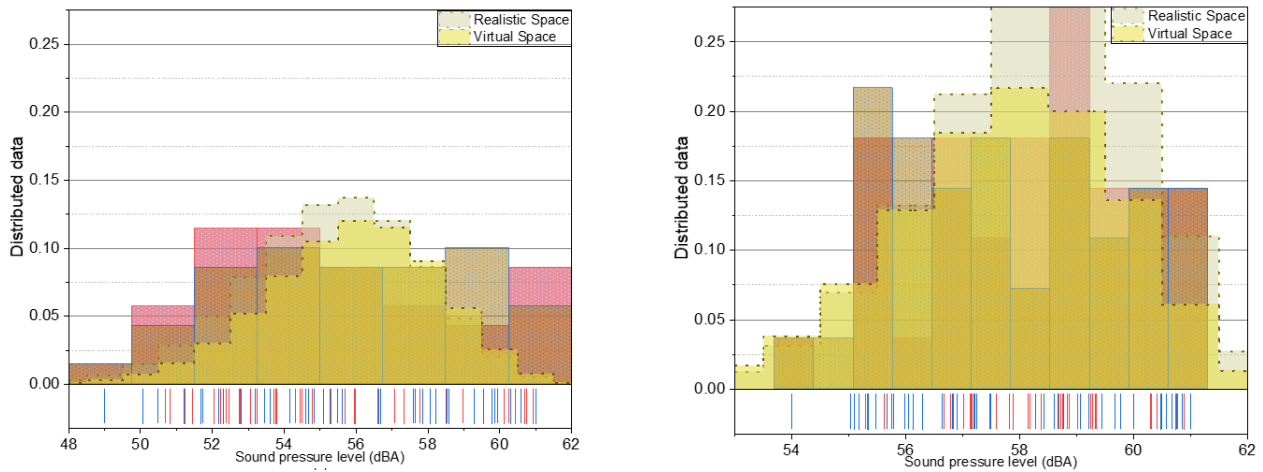


FIGURE 5. Distribution of virtual space and real space under equal sound field.

the audio features of the original speech by the speech feature extraction technique as the multi-step input of the network, decodes the speech data into pinyin sequences by the computation of the GRU network and CTC decoding by the CTC technique, and then performs the process of translation from pinyin to text by the Transformer language model [32].

The Transformer speech model obtains the final text output as the recognition result of the original speech file. Firstly, although the more profound the depth of the neural network, the higher the accuracy of the model, the deeper the depth means, the higher the complexity of the model and the number of GRU layers is set to 3 in this model. Secondly, the category in this experiment is a multi-classification problem, so a suitable Softmax activation function is selected. Thirdly, neural networks require an optimizer to optimize the value of the loss function during training, and the Adam algorithm, which performs well in various environments, is chosen here. Fourth, the learning rate is a measure of the magnitude of the adjustment of the weight parameters; the learning rate is set to 0.0008. The parameter test plots of the network training are shown in Figure 4. Adam is different from the classical stochastic gradient descent method. Stochastic gradient descent maintains a single learning rate (called alpha) for all weight updates, and the learning rate does not change during training. Each network weight (parameter) maintains a learning rate and is adjusted individually as the learning unfolds. The method calculates the adaptive learning rate for different parameters from the budget of the first and second moments of the gradient.

After data processing, there is no significant difference between the reverberation time and spectrum of virtual space and real space on four gradients of 45d BA, 55d BA, 65d BA, and 75d BA. The overall trend of reverberation in virtual and real freedom is the same, and the difference at each frequency is on average 0.0035s higher in virtual space than in an actual room, which can be said to be a slight difference. As shown in Figure 5, the spectrum of sound, the headphones provide

a spectrum in the range below 40Hz, and above 12.5k Hz, and the authentic sound will be different, but this frequency range is not a sensitive area of the human ear. By calculating the difference in sound pressure level in the range of 40Hz-10k Hz, the virtual sound field is 0.012d BA higher than the average difference in sound pressure level of the entire sound field, which also indicates the difference between the two is slight.

Compare the sound distribution of the virtual sound field and the actual sound field. No significant difference appeared in the four good pressure level gradients tested, so the excellent area was compared and analyzed at 55d BA. No significant difference was observed in any of them, and the average difference at 9 points was higher in the actual sound field than in the virtual sound field by 0.04d BA. The average difference at 9 points was higher in the virtual space than in the natural area by 0.02d BA, which illustrated the high accuracy of the calculation of the acoustic engine. The model training process is shown in Figure 6, and for the clarity of the graph, the error word rate is converted into the accuracy rate presented in the chart. The model converges faster, and the accuracy rate tends to be stable above 95% after 200 iteration cycles of the model. The proposed deep, fully convolutional network model can reduce the error rate to 3.23%. The experimental comparison data of the deep, fully convolutional network model can improve speech recognition accuracy. This paper's work will positively affect the understanding and communication of Sichuan and Chongqing cultures.

Too small a learning rate in the training model will lead to slower convergence and produce model training oscillations. In this study, the initial learning rate of the model is set to 0.001; the amount of data per batch will also impact the training effect. Too much data will lead to fewer weight updates and iterations, and too little data may not reach the computational scale of the matrix in training. The number of neurons also has an impact on the feature extraction ability. If there are too many neurons, it will increase the complexity

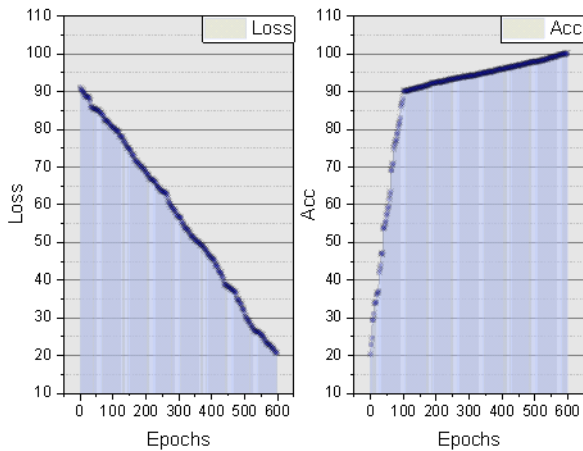


FIGURE 6. Convergence curve of the model training process.

of the network and slow down the learning speed of the network; on the contrary, the network’s learning ability is too poor to obtain the features accurately the choice of the activation function. The activation function introduces a non-linear element to the neural network and plays a vital role in learning the neural network. Emojis have become an essential product of online culture when the network is developing rapidly. The emergence of emoji packs is an excellent way to fill this gap, significantly enhancing the emotional element in the modern network interaction process [33]. It brings friends and relatives closer when interacting based on the network medium; relying solely on text can no longer convey people’s emotional world and cannot get the dynamic transmission brought by language and facial changes during face-to-face conversation. Thus, the combination of emoji packs and dialect has become a hilarious form of entertainment in the era of network communication. This kind of expression combining graphics and words is excellent for avoiding misunderstandings between the two parties in the communication process and creating a beautiful, harmonious, and witty interactive environment.

**B. IMPLEMENTATION OF VISUAL TRANSFORMATION OF DIALECTS IN SICHUAN AND CHONGQING BASED ON VIRTUAL REALITY TECHNOLOGY**

The establishment of an orderly visual flow is achieved through the arrangement of visual elements based on following the habitual visible movement rules of the audience.

By summarizing the research methods after user experience, based on the special characteristics of virtual reality scenes, the interaction design research methods in virtual reality scenes are different from the traditional interaction design research methods. Based on the users’ evaluation of virtual reality interaction, we can derive the users’ views on the product interaction design process in virtual reality scenes under the users’ perspective. Based on users’ evaluation of virtual reality interactions, we can derive users’ perceptions of the interaction design process in virtual reality

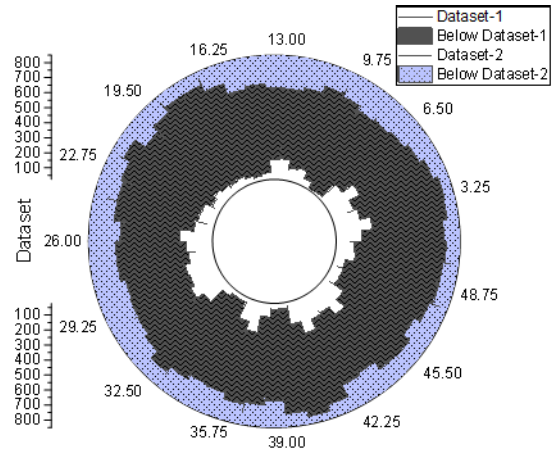
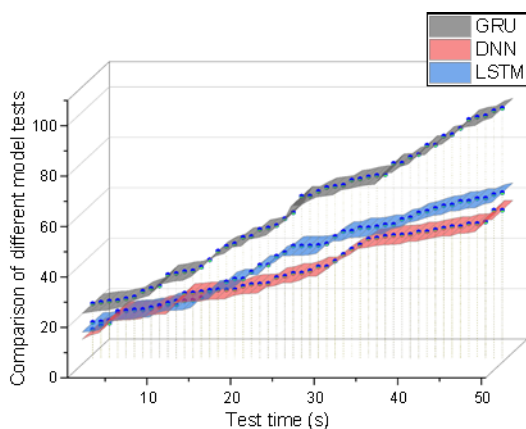


FIGURE 7. Visual conversion accuracy rate.

scenes from their perspectives. 56.17% of users think that the interaction design method in virtual scenes has a positive impact on improving users’ interaction experience, 23.29% of users have a neutral attitude that it has some impact, and the remaining 20.55% of users think it has little impact. Through the analysis, it is concluded that the product interaction design research method based on virtual scenes simplifies the traditional interaction design research process through the combination of virtual and reality, and uses the high immersion of virtual reality scenes to allow users to participate in the interaction design research process, which has a positive effect on improving user experience. We mainly explore the formation of optical flow from the two influencing factors of layout structure and visual hierarchy, firstly, we propose to use the method of grid design in layout design to build a transparent optical system, and secondly, we will use the principle of Gestalt perceptual organization as the theoretical basis to form a clear visual hierarchy through the control of design elements and thus influence the generation of optical flow. The “grid” is composed of horizontal and vertical lines arranged evenly across each other, and grid design is a layout design method based on mathematical proportional relationships.

It is widely used in various design categories of graphic design to assist in coordinating the direct arrangement relationship and to constrain the layout of visual elements in the form [34]. The format of the layout structure through grid design is usually apparent. To establish an orderly optical flow in infographics, it is necessary to form a clear and tidy visual perception of the whole. Then it is inevitable to use grid design methods to organize the visual elements. In the face of a large volume of information, an arbitrary arrangement without the constraints of the grid often puts designers in a dilemma of conflicting visual elements, and the overall effect of the final layout can easily give a haphazard sense and irregularity. The grid sets the layout rules; it can precisely specify the position of visual elements, avoiding disorder and conflict, and the visual elements often have a precise





**FIGURE 8. Virtual reality technology for Sichuan and Chongqing regional dialects Visual transformation Performance test results.**

alignment between them. Similar information is reflected in the layout similarity under the grid's constraints [35], [36]. In addition, due to the mathematical characteristics of the grid design itself, the arrangement of visual elements is by the grid's gradation multiplier, so the grid's layout is easy to form a clear and orderly visible structure. The accuracy of visual transformation is shown in Figure 7 [37].

The data set used for the experiments is divided into a training set containing about 201 minutes of speech data and a test set of 20 minutes in length, and the data in the test set does not intersect with the data in the training set. To better evaluate the model proposed in this paper, this section will be compared with other models of dialect recognition and tested on other dialect recognition models using the same test set to verify the performance test of the method in this paper, as shown in Figure 8. The data that the improved GRU-based recognition method constructed in this chapter has a relative reduction of 2.02% over the DNN model, 0.29% over the LSTM model, and 7.1% over the CTC-based recognition method, which significantly improves the performance of the model, and the proposed recognition method in this paper has some superiority. According to the data of the test set, the proposed recognition method based on GRU acoustic model has better performance in Sichuan and Chongqing dialect speech recognition; even if it can reach the recognition rate of more than 97%, some cases cannot be positively recognized, such as the recognized text has less text than the label or more "um," "ah," etc. The reason for less text may be that the speaker speaks too fast to separate the two pronunciations, while more text may be caused by too much noise in the corpus, so the robustness of the model needs to be improved.

## V. CONCLUSION

In the digital age, Internet information technology has influenced the development of various industries, and education has also been deeply affected and constantly updated its development methods. The combination of virtual reality technology and Sichuan and Chongqing dialects provides

workers with a vast space for practice and research. The visual transformation of dialects can effectively improve communication between dialects in various regions and offer new ideas for promoting the culture of Sichuan and Chongqing cities and preserving Sichuan and Chongqing dialects. We propose improved models and algorithms from two directions: convolutional neural network-based and gated recurrent network-based, establish a detailed model framework from speech pre-processing to final output results, determine the best parameter configuration through multiple debugging, and compare the output results of different models. Through simulation experiments, the convolutional neural network-based model proposed in this study achieves a 3.23% error rate on the Sichuan and Chongqing dialect corpus, while the gated recurrent network-based model achieves a 2.74% error rate on the other hand; both of which are better than the relevant comparison algorithms in the field. Virtual reality technology can meet the experimental conditions of speech talk experiments' visual and auditory aspects. The optical simulation validation experiment and the acoustic simulation validation experiment were conducted. The virtual reality technology simulation degree was verified for the subjective perception of visual scenes, the subjective perception of auditory settings, and objective data, respectively. The experimental results the virtual reality technology is reliable in this study's dialect visual transformation effect.

In this design, there is a problem that the visual translation design of the dialects of Sichuan and Chongqing is not comprehensive enough, and the understanding of the culture of Sichuan and Chongqing is not deep enough. In the future, we should continue to conduct in-depth research on the dialect culture of the Sichuan and Chongqing regions, and search for various possibilities of visual interpretation design for the dialects of the Sichuan and Chongqing regions.

## REFERENCES

- [1] Z. Guo, K. Yu, A. K. Bashir, D. Zhang, Y. D. Al-Otaibi, and M. Guizani, "Deep information fusion-driven POI scheduling for mobile social networks," *IEEE Netw.*, vol. 36, no. 4, pp. 210–216, Jul. 2022.
- [2] Y. Lu, L. Yang, S. X. Yang, Q. Hua, A. K. Sangaiah, T. Guo, and K. Yu, "An intelligent deterministic scheduling method for ultralow latency communication in edge enabled industrial Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 19, no. 2, pp. 1756–1767, Feb. 2023.
- [3] Z. Guo, K. Yu, Z. Lv, K.-K.-R. Choo, P. Shi, and J. J. P. C. Rodrigues, "Deep federated learning enhanced secure POI microservices for cyber-physical systems," *IEEE Wireless Commun.*, vol. 29, no. 2, pp. 22–29, Apr. 2022.
- [4] Y. Li, H. Ma, L. Wang, S. Mao, and G. Wang, "Optimized content caching and user association for edge computing in densely deployed heterogeneous networks," *IEEE Trans. Mobile Comput.*, vol. 21, no. 6, pp. 2130–2142, Jun. 2022.
- [5] Y. He, L. Nie, T. Guo, K. Kaur, M. M. Hassan, and K. Yu, "A NOMA-enabled framework for relay deployment and network optimization in double-layer airborne access VANETs," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 22452–22466, Nov. 2022.
- [6] Y. Yang, "The *dannaku* interface on Bilibili and the recontextualised translation practice: A semiotic technology perspective," *Social Semiotics*, vol. 30, no. 2, pp. 254–273, Mar. 2020.
- [7] Z. Guo, Y. Shen, S. Wan, W.-L. Shang, and K. Yu, "Hybrid intelligence-driven medical image recognition for remote patient diagnosis in Internet of Medical Things," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 12, pp. 5817–5828, Dec. 2022.

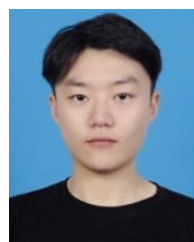
- [8] S. Xia, Z. Yao, Y. Li, and S. Mao, "Online distributed offloading and computing resource management with energy harvesting for heterogeneous MEC-enabled IoT," *IEEE Trans. Wireless Commun.*, vol. 20, no. 10, pp. 6743–6757, Oct. 2021.
- [9] Z. Cai and X. Zheng, "A private and efficient mechanism for data uploading in smart cyber-physical systems," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 2, pp. 766–775, Apr. 2020.
- [10] G. Chen, J. Cui, J. Qian, J. Zhu, L. Zhao, B. Luo, T. Cui, L. Zhong, F. Yang, G. Yang, X. Zhao, Y. Zhou, M. Geng, and J. Sun, "Rapid progress in intelligent radiotherapy and future implementation," *Cancer Invest.*, vol. 40, no. 5, pp. 425–436, May 2022.
- [11] Z. Guo, C. Tang, H. Tang, Y. Fu, and W. Niu, "A novel group recommendation mechanism from the perspective of preference distribution," *IEEE Access*, vol. 6, pp. 5865–5878, 2018, doi: 10.1109/ACCESS.2018.2792427.
- [12] D. Peng, D. He, Y. Li, and Z. Wang, "Integrating terrestrial and satellite multibeam systems toward 6G: Techniques and challenges for interference mitigation," *IEEE Wireless Commun.*, vol. 29, no. 1, pp. 24–31, Feb. 2022.
- [13] K. Zhang, H. Ying, H.-N. Dai, L. Li, Y. Peng, K. Guo, and H. Yu, "Compacting deep neural networks for Internet of Things: Methods and applications," *IEEE Internet Things J.*, vol. 8, no. 15, pp. 11935–11959, Aug. 2021.
- [14] D. Xu, K. Yu, and J. A. Ritcey, "Cross-layer device authentication with quantum encryption for 5G enabled IIoT in industry 4.0," *IEEE Trans. Ind. Informat.*, vol. 18, no. 9, pp. 6368–6378, Sep. 2022.
- [15] F. Ren, "Analysis of the agglomeration of Chinese manufacturing industries and its effect on economic growth in different regions after entering the new normal," *Appl. Math. Nonlinear Sci.*, vol. 6, no. 2, pp. 89–98, Jul. 2021.
- [16] B. Zhu, K. Chi, J. Liu, K. Yu, and S. Mumtaz, "Efficient offloading for minimizing task computation delay of NOMA-based multiaccess edge computing," *IEEE Trans. Commun.*, vol. 70, no. 5, pp. 3186–3203, May 2022.
- [17] Z. Cai, X. Zheng, J. Wang, and Z. He, "Private data trading towards range counting queries in Internet of Things," *IEEE Trans. Mobile Comput.*, early access, Apr. 1, 2022, doi: 10.1109/TMC.2022.3164325.
- [18] Y. Li, J. Zhu, S. C. Hoi, W. Song, Z. Wang, and H. Liu, "Robust estimation of similarity transformation for visual object tracking," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8666–8673.
- [19] S. Kumar, I. D. Popivanov, and R. Vogels, "Transformation of visual representations across ventral stream body-selective patches," *Cerebral Cortex*, vol. 29, no. 1, pp. 215–229, Jan. 2019.
- [20] S. Chota and S. Van Der Stigchel, "Dynamic and flexible transformation and reallocation of visual working memory representations," *Vis. Cognition*, vol. 29, no. 7, pp. 409–415, Aug. 2021.
- [21] J. Sukhera, C. J. Watling, and C. M. Gonzalez, "Implicit bias in health professions: From recognition to transformation," *Academic Med.*, vol. 95, no. 5, pp. 717–723, May 2020.
- [22] Z. Chen, J. Zhang, and D. Tao, "Progressive LiDAR adaptation for road detection," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 3, pp. 693–702, May 2019.
- [23] C. L. Shields, L. A. Dalvin, D. Ancona-Lezama, M. D. Yu, M. Di Nicola, B. K. Williams, J. A. Lucio-Alvarez, S. M. Ang, S. Maloney, R. J. Welch, and J. A. Shields, "Choroidal nevus imaging features in 3,806 cases and risk factors for transformation into melanoma in 2,355 cases: The 2020 Taylor R. Smith and Victor T. Curtin lecture," *Retina*, vol. 39, no. 10, pp. 1840–1851, 2019.
- [24] R. Huda, M. J. Goard, G. N. Pho, and M. Sur, "Neural mechanisms of sensorimotor transformation and action selection," *Eur. J. Neurosci.*, vol. 49, no. 8, pp. 1055–1060, Apr. 2019.
- [25] B. Hatipoglu, C. M. Yilmaz, and C. Kose, "A signal-to-image transformation approach for EEG and MEG signal classification," *Signal, Image Video Process.*, vol. 13, no. 3, pp. 483–490, Apr. 2019.
- [26] R. R. Cong, J. Lei, H. Fu, J. Hou, Q. Huang, and S. Kwong, "Going from RGB to RGBD saliency: A depth-guided transformation model," *IEEE Trans. Cybern.*, vol. 50, no. 8, pp. 3627–3639, Aug. 2020.
- [27] P. Loubeyre, F. Occelli, and P. Dumas, "Synchrotron infrared spectroscopic evidence of the probable transition to metal hydrogen," *Nature*, vol. 577, no. 7792, pp. 631–635, Jan. 2020.
- [28] J. Zhou, J. Yao, W. Zhang, and D. Zhang, "Multi-scale retinex-based adaptive gray-scale transformation method for underwater image enhancement," *Multimedia Tools Appl.*, vol. 81, no. 2, pp. 1811–1831, Jan. 2022.
- [29] S. Anu, K. Muthukumar, M. Punniyamoorthy, S. A. Veerapandian, and G. Sangeetha, "A methodology for the transformation of architectural forms into music and vice-versa for the enhancement of the musical and architectural libraries," *Multimedia Tools Appl.*, vol. 80, no. 7, pp. 10901–10926, Mar. 2021.
- [30] B. H. Yilmaz, C. M. Yilmaz, and C. Kose, "Diversity in a signal-to-image transformation approach for EEG-based motor imagery task classification," *Med. Biol. Eng. Comput.*, vol. 58, no. 2, pp. 443–459, Feb. 2020.
- [31] M. Velez, "'Why take the photo if you didn't want it online?': Agency, transformation, and nonconsensual pornography," *Women's Stud. Commun.*, vol. 42, no. 4, pp. 452–470, 2019.
- [32] M. Labbé and F. Michaud, "RTAB-map as an open-source LiDAR and visual simultaneous localization and mapping library for large-scale and long-term online operation," *J. Field Robot.*, vol. 36, no. 2, pp. 416–446, 2019.
- [33] E. Cerceo, M. Zimmerman, and H. M. DeLisser, "Diversity, equity, and inclusion: Moving from performance to transformation through the arts and humanities," *J. Gen. Internal Med.*, vol. 37, no. 4, pp. 944–946, Mar. 2022.
- [34] S. Erisen, "Incremental transformation of spatial intelligence from smart systems to sensorial infrastructures," *Building Res. Inf.*, vol. 49, no. 1, pp. 113–126, Jan. 2021.
- [35] H. Li, G. Bi, W. Song, and X. Yuan, "Trade credit insurance: Insuring strategy of the retailer and the manufacturer," *Int. J. Prod. Res.*, vol. 60, no. 5, pp. 1478–1499, Mar. 2022.
- [36] C. Yu, J. Liu, J. Zhang, K. Xue, S. Zhang, J. Liao, Q. Tai, and D. Zhu, "Design and optimization and experimental verification of a segmented double-helix blade roller for straw returning cultivators," *J. Chin. Inst. Eng.*, vol. 44, no. 4, pp. 379–387, May 2021.
- [37] J. Wang, Y. Hou, L. Jiang, and L. Zhang, "Robust stability and stabilization of 2D positive system employing saturation," *Circuits, Syst., Signal Process.*, vol. 40, no. 3, pp. 1183–1206, Mar. 2021.



**JIANCHAO ZHAO** was born in 1975. He received the master's degree from the Beijing University of Posts and Telecommunications, in 2005. Currently, he is an Associate Professor with the Henan Polytechnic Institute. His main research interests include AI and information security.



**ZHIZHI FENG** was born in Sichuan, China, in 1986. He received the master's degree in visual communication design from Sichuan Fine Arts Institute, China. Currently, he is with the School of Humanities and Arts, Chongqing University of Science and Technology. His research interests include visual communication design, digital arts, and intangible cultural arts.



**ZHENG DU** was born in Chongqing, China. He received the master's degree in computer science from Imperial College London. Currently, he is working with Baidu Online Network Technology Company—a big famous AI company, where he participates in the field of autonomous driving. His research interests include voice interaction, data analysis, electronic art, and photography.