IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 January 2024, accepted 12 February 2024, date of publication 22 February 2024, date of current version 7 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3369036

RESEARCH ARTICLE

Multi-View Video Quality Enhancement Method Based on Multi-Scale Fusion Convolutional Neural Network and Visual Saliency

WEIZHE WANG AND ERZHUANG DAI
College of Modern Information Technology, Henan Polytechnic University, Zhengzhou 450018, China
Corresponding author: Weizhe Wang (25011@hnzj.edu.cn)

ABSTRACT This study aims to improve the quality of multi-view video, and designs a new method by integrating Convolutional neural network (CNN), visual saliency detection and image enhancement theory. The experimental results show that the proposed visual saliency detection model and convolution filter sensor have made remarkable progress. The superiority of the visual saliency detection model is helpful to accurately locate the key features of the image and provide accurate targets for subsequent enhancement processing. The convolution filter sensor improves the peak value of the image, narrows the gap with the original image and improves the visual effect. Supplementary experiments further verify the effectiveness of the method. Through the quantitative comparison between SSIM and MS-SSIM, the method is obviously superior to the existing methods on several data sets, showing a robust video quality enhancement effect. These results highlight the superiority and robustness of the method, and bring strong empirical support to the field of multi-view video quality enhancement, which is expected to have an important impact in practical applications.

INDEX TERMS Convolutional neural network, visual salience, multi-view video, quality enhancement, multiscale, image detection.

I. INTRODUCTION

With the development of the times, network technology is becoming increasingly mature and video technology is also making continuous progress. People can often find video equipment in every corner around them. The video system has really integrated into people's lives, whether at the crossroads of traffic or in the residential area of daily life [1]. The video system includes a video surveillance system, video conference system, and other types of systems, these diverse systems constitute a complete video system ecology, thus providing important support for the development of information transmission through video. However, the equipment required for video transmission varies from professional video cameras to small and portable USB cameras and even mobile phone cameras, all of which have different sizes and functions, but all face adverse effects of environmental factors on the quality of video collection [2]. Generally speaking, due to equipment accuracy, light, or other reasons, the acquired video images usually have image quality problems [3]. The common reasons for this problem are the lossy encoding and decoding process, low-resolution acquisition device, relative movement of image acquisition device, too dark field environment, or strong highlight environment, etc. Images with quality loss often have problems such as low brightness, obscure detail information, excessive noise, low contrast, etc., which makes it difficult for observers to obtain useful information from these videos. Therefore, many studies have analyzed these problems.

Fan et al. [4] pointed out that the current 3D multi-view video technology is not mature, and many aspects need to be improved, such as the drawing technology based on the depth image. The depth image obtained by the existing

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

depth estimation algorithm has some problems, such as occlusion, mismatching, and void, and the depth video obtained by the existing depth video acquisition system has low resolution. Therefore, it needs to be enhanced by pre-processing, post-processing, or depth image reconstruction to obtain satisfactory rendering video [4]. Turkoglu et al. [5] proposed that to adapt to different display sizes, video is required to adapt to different aspect ratios. However, the video resolution of mobile terminals is low, the quality is poor, and there are defects. This requires the relocation of the original 3D multi-view video so that the processed video can not only maintain a high quality but also adapt to displays of different sizes [5]. Montero-Plata and Pruvost-Delaspre [6] argued that the study of visual saliency has always been an important aspect of the field of human vision research. And the saliency detection model proposed by visual saliency can be used for target tracking, image segmentation, relocation of images or videos, and other fields. Due to the influence of the acquisition system, storage medium, compression coding, and transmission equipment, noise or interference are inevitably introduced in the collection, storage, processing, and transmission of video, resulting in the decline of video quality [6]. Dewi et al. [7] pointed out that due to the accuracy of equipment in the shooting process, compression format and algorithm in the preservation process, noise interference in the transmission process, distortion caused by compression technology, and others, all these factors will lead to the difference between images and videos after processing and transmission and the original images seen by the photographer. Sometimes there is even serious distortion [7]. Ding et al. [8] put forward that Three Dimensional Television (3DTV) system usually uses multiple cameras to obtain scene information from different viewpoints, and the amount of data transferred is very large, which is not suitable for practical video applications [8]. Wei et al. [9] proposed that Depth Image Based Rendering (DIBR) technology could produce an interactive 3D visual experience by generating virtual viewpoint information at the client end. Because of the depth discontinuity in the depth image, new exposed areas (i.e., new voids) appear in the virtual rendering image. At present, some deep preprocessing techniques have been proposed to deal with the void problem [9]. Phaphuangwittayakul et al. [10] also proposed a lot of post-processing technologies for the encoding distortion of depth images, and high-resolution depth images can be obtained through reconstruction for low-resolution depth images. The enhanced depth image can be acquired through processing and a 3D video of higher quality can be obtained through drawing. Therefore, some effective processing of the depth images can well enhance the 3D video [10].

On this basis, combining video image enhancement theory and visual saliency detection theory, a multi-view video quality enhancement method based on the Convolution Neural Network (CNN) is proposed. Moreover, simulation experiments are performed using various databases

II. RELEVANT THEORIES AND METHODS

A. CONVOLUTIONAL NEURAL NETWORK

One of the representative deep learning (DL) methods is CNN, a type of feedforward neural network with a deep structure and convolution calculation [11]. The Shift-Invariant Artificial Neural Network (SIANN) is another name for CNN, which can categorize input data with translation invariance in accordance with its hierarchical structure. The convolution, pooling, and fully connected (FC) layers make up the foundation of CNN. The middle layer of the LeNet architecture is made up of two convolution layers, two down sampling layers, and two complete connection layers in addition to the input layer and the output layer. The CNN is used to classify the handwritten grayscale image, and the results can be calculated and output [12]. The network structure is displayed in Figure 1.

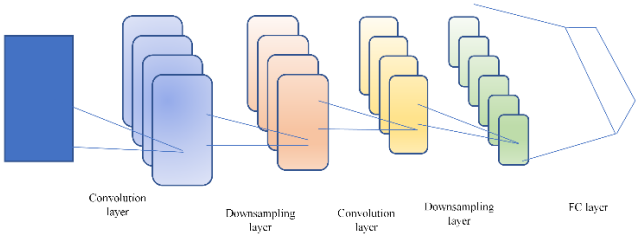

FIGURE 1. LeNet structure diagram.

Since then, CNN has largely adopted the LeNet architecture, which consists of three main components: a convolution layer, a pooling layer, and a FC layer. Feature graphs are obtained through convolution and pooling operations, and they are then transformed into one-dimensional vectors and input into the FC layer, where two or more classifications are achieved through the classification layer [13]. This section provides a brief explanation of the principle and purpose of each CNN component, using the structure of LeNet as an example [14]. The image of the particular function is presented in Figure 2.

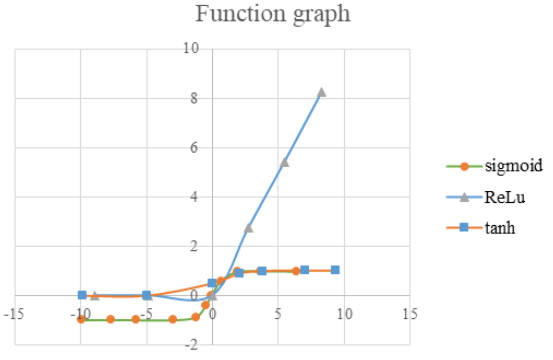

FIGURE 2. Common activation function diagram.

The operation process of uniform pooling and maximum pooling with a pool window of 22 and a step size of 2 is shown in Figure 3 [15].

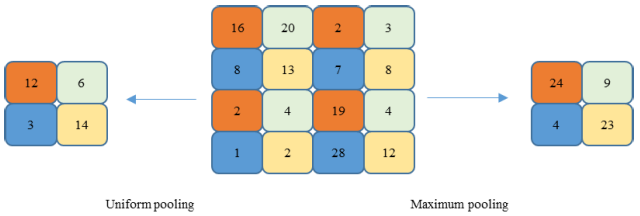

FIGURE 3. Schematic diagram of pool operation.

Only the convolution results are organized into column vectors, and the entire connection layer only requires a modest amount of processing [16]. However, compared to the convolution layer, the FC layer's parameters account for a significant amount because of the full connection mode [17]. Figure 4 illustrates the FC layer's underlying theory.

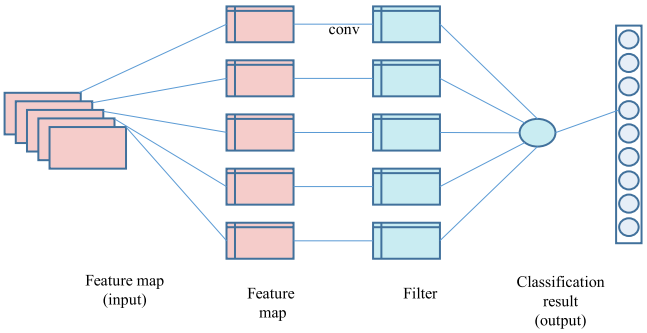

FIGURE 4. Schematic diagram of the FC layer.

In Figure 4, taking VGG19 as an example, only the parameters of the three FC layers at the end of the network account for 90% of the parameters of the whole network model. Therefore, in CNN, the focus of computing acceleration and optimization is convolution layer, while the emphasis of parameter optimization and weight cutting lies in the FC layer [18].

B. IMAGE QUALITY EVALUATION

A crucial component of improving video image quality is image quality evaluation. As mentioned earlier, obtaining multi-view video images with rich details and high spatial and spectral resolution is the goal of integrating CNN and visual saliency. This will assure the accuracy and reliability of later computer interpretation. The evaluation of the video enhancement method is only indirectly done through the evaluation of multi-view video images because the image fusion algorithm itself determines the quality of multi-view video images. Currently, subjective and objective evaluations are utilized to determine how well an image is made [19].

1) SUBJECTIVE EVALUATION

To evaluate and score the quality of the fusion image, relevant personnel must first see the fusion image to be evaluated and compare it with the scoring standard. This is how subjective evaluation is often carried out using the visual observation system of human eyes. Table 1 exhibits the current general subjective evaluation criteria and associated scores.

TABLE 1. Subjective evaluation scoring.

Score	Quality scale	Obstacle scale
5	Very good	No deterioration of image quality can be seen
4	Good	It shows that the image quality has deteriorated, but it does not hinder the viewing
3	General	It clearly shows that the image quality has deteriorated, which slightly hinders viewing
2	Bad	Be detrimental to the viewing
1	Very bad	Very seriously hinder viewing

Table 1 describes that the evaluation criteria are grouped into five grades, and the blockage scale and quality scale can be used to swiftly assess the image quality. Actually, if conditions permit, a predetermined number of individuals can be chosen to rate the image under consideration, and the average value can then be employed to determine the subjective evaluation outcome of the image.

2) OBJECTIVE EVALUATION

The assessment index of photographs is adopted to conduct an objective evaluation. Image assessment indexes are certain numerical values based on the pixel values of images and coupled with various calculation methods. They have distinct physical meanings. These indexes can be used as a reflection of the actual image quality and have the advantages of objectivity, comprehensiveness, and efficiency. The objective evaluation has not yet developed into a single standard. To fulfill the goal of image evaluation, appropriate indexes are often chosen based on the situation. Several image evaluation indexes' physical definitions and computation methods are described below [20].

a: PEAK SIGNAL-TO-NOISE RATIO (PSNR)

One statistic that can be used to determine how much the fused image has been warped during the fusion process is the PSNR, or effective SNR. It is assumed that the PSNR of image x and reference image r is:

$$PSNR_{X,R} = 10log\frac{k^2}{\frac{1}{M\times N}\sum_{i=1}^{M}\sum_{j=1}^{N}(X(i,j)-R(i,j))^2} \tag{1}$$

$M\times N$ refers to the image size; (i,j) represents the pixel position; K refers to the maximum gray level of the image.

b: STRUCTURAL SIMILARITY (SSIM)

SSIM is used to measure the structural similarity of two images. The SSIM between image X and reference image R is defined as:

$$SSIM_{X,R} = \sum_{x,r}\frac{2\mu_x\mu_r + C_1}{\mu_x^2\mu_r^2 + C_1}\cdot\frac{2\sigma_x\sigma_r + C_2}{\sigma_x^2\sigma_r^2 + C_2}\cdot\frac{\sigma_{xr} + C_3}{\sigma_x\sigma_r + C_3} \tag{2}$$

x,r represent the image blocks of images X and R; $\mu_x^2\mu_r^2$ refers to the mean; $\sigma_x^2\sigma_r^2$ indicates the standard deviation; σ_{xr} means the covariance; C_1, C_2 and C_3 are the stability constants of the algorithm.

c: CORRELATION COEFFICIENT (CC)

CC is employed to measure the degree of linear correlation between images. The linear CC between images X and R can be written as equation (3):

$$CC = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(X(i,j)-\bar{X})(R(i,j)-\mu)}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}\left(X(i,j)-\bar{X}\right)^2\left(\sum_{i=1}^{M}\sum_{j=1}^{N}(R(i,j)-\mu)^2\right)}} \tag{3}$$

X expresses the average pixel value of the image; (i,j) represents the pixel position; μ signifies the average pixel value of the image R.

d: SPATIAL FREQUENCY (SF)

Image quality is measured based on the distribution of horizontal gradient and vertical gradient (also called spatial row frequency and column frequency) of SF image. The SF of image X reads:

$$SF = \sqrt{RF^2 + CF^2} \tag{4}$$

RF is as follows:

$$RF = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(X(i,j)-X(i,j-1))^2} \tag{5}$$

The expression of CF is indicated in equation (6):

$$CF = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(X(i,j)-X(i-1,j))^2} \tag{6}$$

e: STANDARD DEVIATION (STD)

Std measures the contrast and dispersion of images based on the concept of statistics. The Std of image X is as follows:

$$Std = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(X(i,j)-\mu)^2} \tag{7}$$

μ stands for the average pixel value of image X; $M\times N$ and (i,j) are the image size and pixel position, respectively.

f: AVERAGE GRADIENT (AVG)

AvG reflects the gradient distribution of the whole image, and the AvG of image x is illustrated in equation (8):

$$AvG\frac{1}{MN}\sum_{i-1}^{M}\sum_{j=1}^{N}\sqrt{\frac{\nabla X_x^2(i,j)+\nabla X_y^2(i,j)}{2}} \tag{8}$$

In equations (9) and (10):

$$\nabla X_x^2(i,j) = X(i,j) - X(i+1,j) \tag{9}$$

$$\nabla X_y^2(i,j) = X(i,j) - X(i,j+1) \tag{10}$$

∇ means gradient operation; $M\times N$ and (i,j) denote image size and pixel position, respectively.

III. RESEARCH METHODS AND DESIGN

A. VISUAL SALIENCY MODEL

To reflect the specific performance and function of the design model in this study, part of the video is used as the research object to test and evaluate the model. During the study, each frame of the video is divided into 8 8 image blocks, and the DCT coefficient can reflect the energy of each block. In the DCT transform, Direct Current (DC) coefficient represents the average energy of all pixels in the image block, while the Alternating Current (AC) coefficient represents the detailed frequency features of the image block [21]. Since the Cr and Cb components in the image mainly include color information, the AC coefficient of the Y component is employed to represent the texture information of the image. In z-scan, the first 9 AC coefficients can represent most of the energy of frequency information, so the first 9 low-frequency AC coefficients are used to express the texture features of image blocks. At the same time, salient regions attract attention mainly because of their contrast with the features of surrounding areas. Thus, the direct method to extract salient regions is to calculate the feature contrast between image blocks. The human vision system is more sensitive to the feature comparison between near image blocks than the feature comparison between distant image blocks. Therefore, using this feature, the Gaussian model of spatial distance is adopted to calculate the weight of image block feature comparison [22]. The saliency of image block i can be expressed as equation (11).

$$F(I_i) = \sum_{j\neq i}\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-1_{ij}^2}{2\sigma^2}}U_{ij} \tag{11}$$

I_i refers to the spatial distance between image blocks i and j, and U_{ij} means the comparison of texture features of the two image blocks. The expression of Uij can be written as equation (12):

$$U_{ij} = \frac{\sqrt{\sum_i\left(B_j^i - B_j^i\right)^2}}{\sum_i\left(B_j^i + B_j^i\right)} \tag{12}$$

b represents texture features.

For video saliency maps, the detection of motion information is an important content of saliency extraction. In this study, motion vectors extracted from video streams are used to detect motion regions. However, there are many problems in detecting moving objects with motion vectors. An eight-parameter global motion model is used to estimate the motion of the camera and the background, and then a more accurate motion vector is obtained by motion compensation. Assuming that (x, y) is a point coordinate of the current frame, the corresponding point coordinate (x′,y′) in the reference frame can be obtained according to the 8-parameter global motion model. The calculation equation is as follows:

$$x^i = \frac{m_0x + m_1y + m_2}{m_6x + m_7y + 1},\ y^i = \frac{m_3x + m_4y + m_5}{m_6x + m_7y + 1} \tag{13}$$

m represents eight parameters of the model.

In this study, the uncertainty in saliency detection is estimated to improve the accuracy of saliency detection. The uncertainty of this study is mainly calculated from two aspects. 1) The closer the spatial distance is to the concentrated salient area, the more likely it is to become a salient point. Assuming that G is the original true saliency map of the image, the center position of the saliency map is calculated with equations (14) and (15):

$$x_c = \frac{1}{M}\sum_{(x,y)\in R_S} xG_{x,y} \tag{14}$$

$$y_c = \frac{1}{M}\sum_{(x,y)\in R_S} yG_{x,y} \tag{15}$$

R_S represents the set of all salient pixels of the true salient graph, and M indicates the total number of pixels in R_S [23].

B. MULTI-VIEW VIDEO ENHANCEMENT TECHNOLOGY

One essential component of 3D video systems is rendering based on depth images. However, the performance of the next virtual viewpoint rendering will be significantly impacted by the distortion of depth coding. In this study, a virtual viewpoint rendering method for depth images based on CNN is proposed. In this method, the depth value of each pixel is first estimated using the least mean square (LMS) approximation polynomial technique, followed by the estimation of the filtering window of each pixel in the depth image, and finally the estimated depth image is filtered using the weighted pattern filter to produce the depth filtered image. Figure 5 plots the structure block diagram of the method:

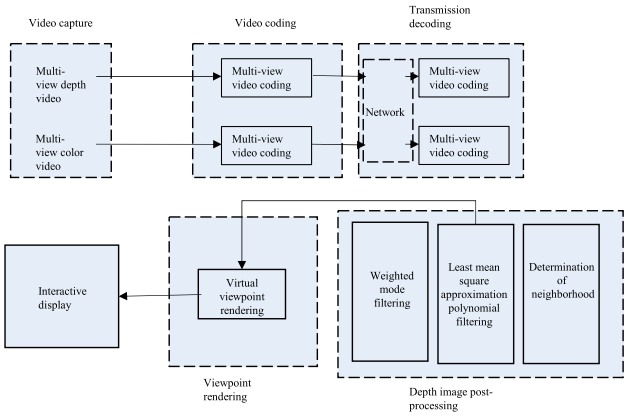

FIGURE 5. Schematic diagram of multi-view video enhancement method.

Figure 5 details that in traditional filtering methods, the filtering window usually adopts a fixed block size. However, the fixed filtering window does not fully consider the structural information of the image. For example, the structure of the edge area of the image is irregular, and if the fixed block size is used for filtering, the edge will be blurred. To solve the above problems, this study proposes a method of establishing filter window. By establishing the optimal size in different directions, an adaptive filter window is constructed. In this study, the intersection of confidence intervals is utilized to determine the optimal scale. Because color images have rich texture information, polynomial kernel function is used to convolve color images:

$$\gamma_{h_j,\theta_k}(p) = (\tilde{I}_{t,i} * g_{h_j,\theta_k})(p) \tag{16}$$

γ_{h_j,θ_k} is the estimated value of the h_j scale of the current pixel in θ_k direction. Then, the estimated values of different scales in various directions are employed to calculate the intersection points of confidence intervals. Finally, the maximum scale value with non-empty intersection points in all scales is selected as the optimal scale of the current pixel in θ_k direction, and the optimal scales in different directions are connected to form the filtering window of the current pixel.

The compressed depth map can be regarded as adding a noisy image to the original depth map, and the LMS polynomial smoothing filter is a method of fitting a low-order polynomial with a noisy signal.

In this study, Collaborative Research in Computational Neuroscience (CRCNS) database is used to verify the feasibility of the proposed saliency model [24]. The eye movement data is recorded using an eye tracker test on eight observers. Two evaluation indexes are employed to compare the significance model. Normalized Scan Path Salience (NSS) is used to measure the degree of agreement between the focus area of human eyes and the saliency map of the model. Through normalization of the saliency map, its mean is zero and it has unit Std. Then, according to the corresponding eye movement data, the focus area of human eyes corresponds to the points in the saliency map, and the corresponding saliency of these points is averaged, and the obtained values are called NSS. The second evaluation index is the area under the Receiver Operating Characteristics Curve (OC-ROC). In this evaluation index, the saliency map is threshed with different thresholds to obtain the ROC curve, and then the area under the ROC curve is calculated, that is, the OC-ROC. The code base of the designed CNN model algorithm is http://tecdat.cn/?p=18149.

To sum up, the pre-trained CNN (each branch contains a series of convolution and pooling layers) is used to capture the feature information at different scales. These branches work in parallel in the architecture to ensure the full use of multi-scale information. As a basic feature extractor, it is used to extract preliminary features from multi-view videos. Meanwhile, the feature maps from different branches are fused, and the multi-scale feature maps are combined into a higher-level feature representation through convolution and up-sampling operations. Meanwhile, the design of convolution filter sensor is to improve image quality and keep image details. The design principle of the sensor is to optimize the filter parameters to ensure that the image quality is enhanced without introducing adverse effects, such as blur or distortion.

IV. RESULTS AND DISCUSSION

A. FEASIBILITY ANALYSIS OF THE SALIENCY MODEL

To highlight the advantages and practicability of the designed model, the designed model is compared with the saliency models of the two time-space zones during the research

process to test the effectiveness of the designed model. (1) Bayesian Surprisemodel (BS); (2) Self-Resemblance model (SR). BS model mainly uses scintillation and motion features to obtain temporal significance and adopts color, density, and edge features to detect spatial significance. SR model principally uses a local regression kernel to test the similarity between each pixel and its surrounding pixels. The specific results are revealed in Figure 6:

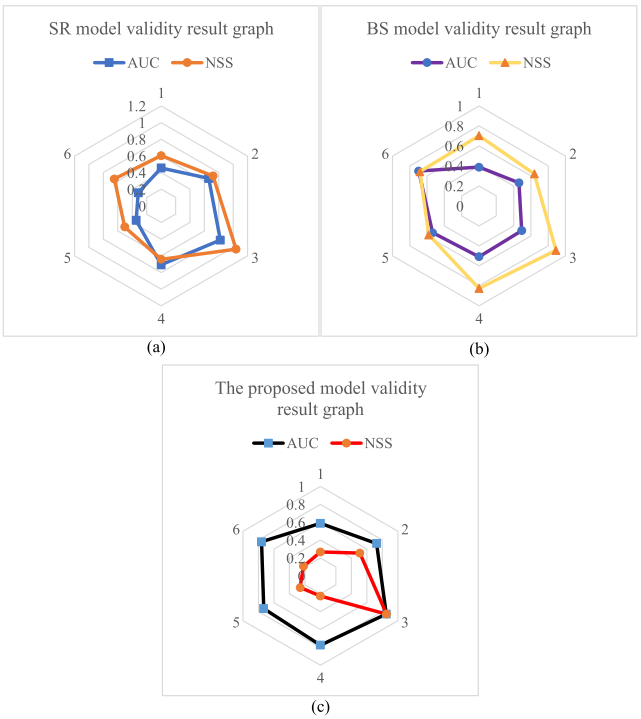

FIGURE 6. Results of effectiveness of multi-view video enhancement method. (a) Results of SR model; (b) Results of BS model; (c) Study and design the model result diagram.

Figure 6 demonstrates that in six groups of comparative data experiments, the AUC value of the visual saliency model researched and designed here is larger than that of SR model and BS model in comparative experiments. Furthermore, the NSS of the visual saliency model researched and designed is smaller than that of other comparative model experimental groups. Therefore, there is a high degree of coincidence between the focus area of human eyes and the saliency map of the model, which shows that the designed visual saliency method is superior to other methods.

B. CNN FILTERING

In this method, three-dimensional video sequences ''Ballet'' and ''Breakdancers'' provided by Microsoft are used for experiments. In the experiment, the color video sequence and depth video sequence of the fourth viewpoint are selected, and the length of each sequence is 50 frames. To illustrate the effectiveness of this method, it is compared with two commonly used depth filtering methods: joint bilateral filtering and trilateral filtering. The two Gaussian kernels of the joint bilateral filter are 15 and 0.025 respectively, the three Gaussian kernels of the trilateral filter are 15, 0.025, and 0.025 respectively, and the filtering windows are all 5×5. The experimental results are plotted in Figure 7:

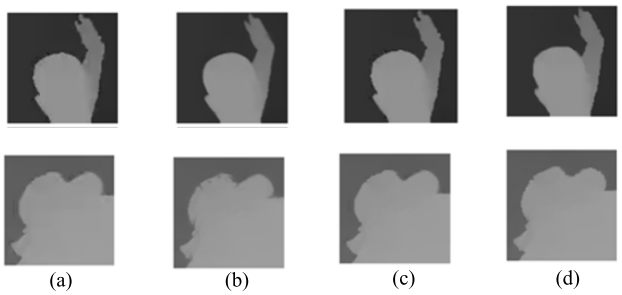

FIGURE 7. Image filtering result diagram. (a) decoding processing result diagram; (b) Figure of bilateral filtering results; (c) Trilateral filtering result; (d) Convolution filtering result diagram.

In Figure 7, the local enlarged images of ''Ballet'' and ''Breakdancers'' depth images processed by different methods are given. Figure 7 shows that the depth images processed by this method can basically eliminate the coding noise information, and produce satisfactory sharp edges and smooth contours. Figure 7 also presents the corresponding local enlarged picture of the drawn image drawn by the processed depth image. The drawn image drawn by the depth image processed by this method can keep the contour information of the object well, and obviously reduce the mapping phenomenon of the background covering the foreground caused by the depth distortion, thus improving the quality of the virtual viewpoint image. The PSNR values between images are further analyzed and the results are suggested in Figure 8:

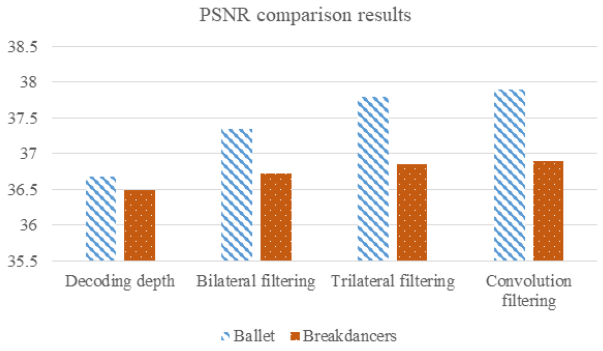

FIGURE 8. Result of image filtering PSNR comparison.

The average PNSR comparison findings between the original virtual viewpoint images and the produced depth images processed using various approaches are revealed in Figure 8. The photos demonstrate how the convolution filtering technique proposed in this study can raise the images' PSNR values. The results of this method are clearly superior to those of other ways since the narrower the gap between the photos and the original images, the higher the PSNR value must be.

C. EVALUATION OF VIDEO QUALITY ENHANCEMENT METHODS

To comprehensively highlight the performance of the designed model, the used CNN technology is trained and tested through 1000 samples. Then the designed video quality enhancement method is tested and evaluated, in which the designed model is compared with the Moving Pictures Experts Group (MPEG) image processing method. MPEG is the international standard of moving pictures compression algorithm, which has been supported by almost all computer platforms. It is a very popular image processing method. Based on this, the encoding and decoding speed and image processing quality of the designed model are evaluated, so as to comprehensively evaluate the performance of the designed model. The results of the comparative test between the designed model and the MPEG method are demonstrated in Figure 9.

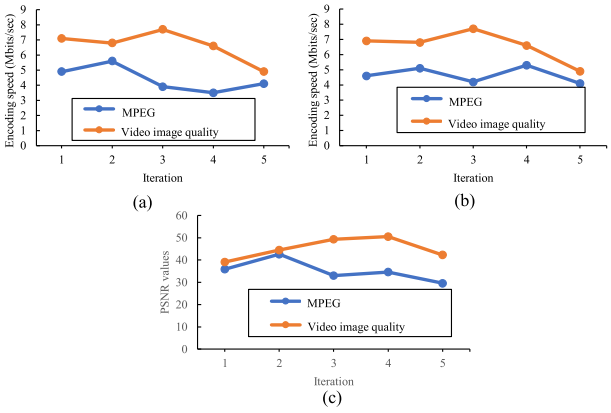

FIGURE 9. Performance evaluation results of video quality enhancement method. a: encoding speed; b: decoding speed; c: image processing effect.

In Figure 9, after evaluation, it is found that the encoding and decoding speed of the MPEG method is around 4-6Mbits/sec, while that of the designed model is around 5-8Mbits/sec. The PSNR value of image processing of the MPEG method and the designed model is about 30-44 and 40-53. It means that the designed model has better image processing performance than the MPEG so that the image can obtain better quality after processing.

Meanwhile, three specific public datasets are further selected, including: MCL-JCV dataset(https://mcl.usc.edu/mcl-jcv-dataset/) and MVPSynth dataset (https://phuang17.github.io/DeepMVS/mvs-synth.html). These datasets contain multi-view videos captured in real environment, covering indoor and outdoor scenes, and are suitable for evaluating video quality from multi-camera perspectives. According to the above datasets, the effect of the multi-view video quality enhancement method is studied and designed. Taking Structural Similarity (SSIM) and MS-SSIM as evaluation indexes, the proposed multi-view video quality enhancement method is compared with three methods, namely, deep learning method, multimodal fusion method and perception-driven image enhancement method. The same dataset and parameter settings are used to make a fair comparison. The specific experimental results are shown in Figure 10:

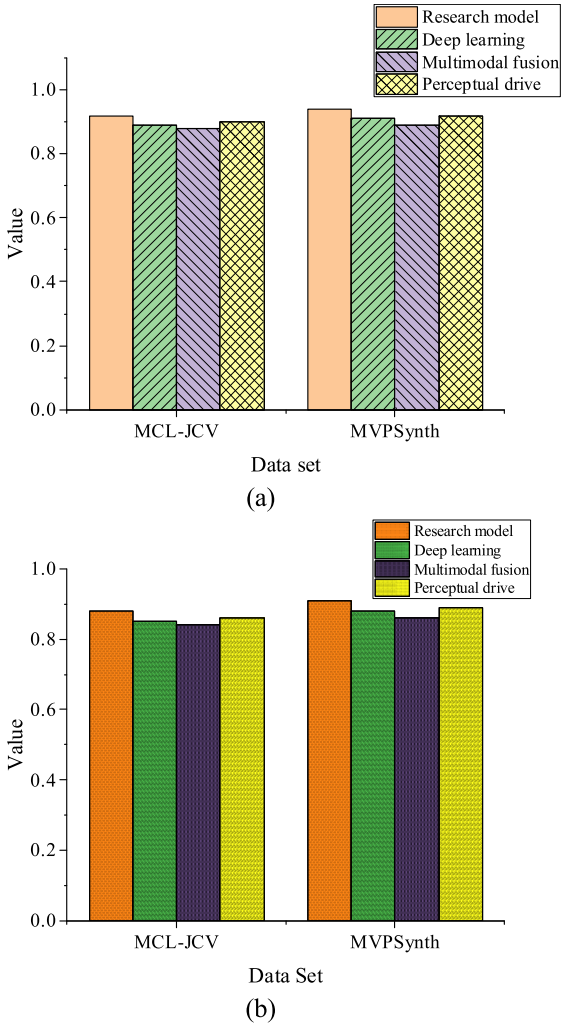

FIGURE 10. Research model comparative analysis result graph a. SSIM Result map b. MS-SSIM Result map.

Note: The closer the scores of SSIM and MS-SSIM are to 1, the better the image quality is.

In Figure 10, according to the evaluation results of SSIM and MS-SSIM, the proposed method shows higher structural similarity and multi-scale structural similarity in multi-view video quality enhancement, and shows better image quality improvement effect compared with the contrast method. These results verify the effectiveness and superiority of the proposed method, and provide strong theoretical support for the field of multi-view video quality enhancement.

D. DISCUSSION

As the importance of messaging efforts continues to grow, video as the main form of efficient and intuitive information transmission has become the main method of information transmission. However, there are often many problems in the current video transmission, among which the more

prominent problem is the loss of video quality, resulting in video information transmission is not clear and explicit. Thus, to effectively improve the quality of video transmission and solve the problem of its quality loss, this study designs the use of CNN technology for video processing to improve the quality of video transmission and optimize the transmission of video information. Here, a CNN video processing model is designed, and its performance in the actual video processing process is tested and evaluated. The results reveal that the AUC value of the designed visual saliency model is greater than that of the SR model and BS model, and the NSS of the designed model is smaller than that of the models of the experimental group. Thereby, there is a high degree of coincidence between the focus area of human eyes and the saliency map of the model. It can be concluded that the designed visual saliency method is superior to other methods.

In addition, the two Gaussian kernels of the joint bilateral filter selected here are 15 and 0.025, and the three Gaussian kernels of the trilateral filter are 15, 0.025, and 0.025 respectively. The filtering Windows are all 5×5, and further experimental tests are carried out. The result shows the comparison of average PNSR between the depth image drawn by different methods and the original virtual viewpoint image. The photos demonstrate that the proposed convolution filtering technique can raise the images' PSNR values. The larger the PSNR value, the smaller the gap between the picture and the original photos. The results of the designed method are obviously better than those of other methods. Meanwhile, compared with the study of Webber et al. [25], this study adopts a more advanced technical model to deal with video processing in a more comprehensive way. Meanwhile, the video evaluation index is more comprehensive, which can effectively highlight the overall performance of the designed model.

V. CONCLUSION

By integrating CNN and visual salience theory, this study promotes the progress of multi-view video technology. Using 3D video enhancement technology, the quality of multi-view video is effectively improved. Firstly, the saliency detection model of multi-view video is discussed from the perspective of human attention, and the method of 3D video enhancement is deeply studied. These studies fully combine the perceptual characteristics of visual salience. The significance detection model based on CNN is combined with the way to improve the quality of multi-view video. Through simulation and contrast experiments, the following accurate results have been obtained on several video datasets. The research results show that the explored and developed visual saliency method is obviously superior to other methods by comparing two evaluation indexes of saliency maps obtained by different methods. Meanwhile, the convolution filtering model can improve the peak value of the image and approximate the original image. The designed model can not only be used as theoretical guidance, but also help to improve the quality of subsequent experiments. The shortcomings of this study are that the sample size is insufficient and the time is limited, and there are some defects in depth and breadth. Although the sample size is small, the experimental results are generally applicable. The research scope will be expanded in the future. The research of deep learning network will keep up with the trend of the times, new technologies will be updated and applied to the follow-up research, and theory and practice will be fully combined for in-depth exploration.

DATA AVAILABILITY STATEMENT

The data used to support the findings of this study are included within the article and its supplementary material. If someone wants to request the data from this study, please contact the corresponding author. (Weizhe Wang, e-mail: 25011@hnzj.edu.cn)

CONFLICT OF INTEREST(COI) STATEMENT

The authors have no conflict of interest.

REFERENCES

[1] B. Jing, H. Ding, Z. Yang, B. Li, and Q. Liu, "Image generation step by step: Animation generation-image translation," *Appl. Intell.*, vol. 52, pp. 8087–8100, Oct. 2022.

[2] Z. Cui, Y. Ito, K. Nakano, and A. Kasagi, "Anime-style image generation using GAN," *Bull. Netw., Comput., Syst., Softw.*, vol. 11, no. 1, pp. 18–24, 2022.

[3] T. M. Ghazal, "Convolutional neural network based intelligent handwritten document recognition," *Comput., Mater. Continua*, vol. 70, no. 3, pp. 4563–4581, 2022.

[4] X. Fan, W. Zhang, C. Zhang, A. Chen, and F. An, "SOC estimation of Li-ion battery using convolutional neural network with U-Net architecture," *Energy*, vol. 256, Oct. 2022, Art. no. 124612.

[5] M. Turkoglu, D. Hanbay, and A. Sengur, "Multi-model LSTM-based convolutional neural networks for detection of apple diseases and pests," *J. Ambient Intell. Humanized Comput.*, vol. 13, no. 7, pp. 3335–3345, 2022.

[6] L. Montero-Plata and M. Pruvost-Delaspre, "Shaping the anime industry: Second generation pioneers and the emergence of the studio system," in *A Companion to Japanese Cinema*, vol. 23, 2022, pp. 215–246.

[7] C. Dewi, R. C. Chen, Y. T. Liu, and H. Yu, "Various generative adversarial networks model for synthetic prohibitory sign image generation," *Appl. Sci.*, vol. 11, no. 7, p. 2913, 2021.

[8] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang, and J. Tang, "CogView: Mastering text-to-image generation via transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 19822–19835.

[9] S. J. Wei, Y. H. Chen, Z. R. Zhou, and G. L. Long, "A quantum convolutional neural network on NISQ devices," *AAPPS Bull.*, vol. 32, no. 1, pp. 1–11, 2022.

[10] A. Phaphuangwittayakul, Y. Guo, and F. Ying, "Fast adaptive meta-learning for few-shot image generation," *IEEE Trans. Multimedia*, vol. 24, pp. 2205–2217, 2022.

[11] M. Torres and F. Cantú, "Learning to see: Convolutional neural networks for the analysis of social science data," *Political Anal.*, vol. 30, no. 1, pp. 113–131. 2022.

[12] P. R. A. S. Bassi and R. Attux, "A deep convolutional neural network for COVID-19 detection using chest X-rays," *Res. Biomed. Eng.*, vol. 38, no. 1, pp. 139–148, 2022.

[13] S. Na, M. Do, K. Yu, and J. Kim, "Realistic image generation from text by using BERT-based embedding," *Electronics*, vol. 11, no. 5, p. 764, 2022.

[14] T. Schmale, M. Reh, and M. Gärttner, "Efficient quantum state tomography with convolutional neural networks," *npj Quantum Inf.*, vol. 8, no. 1, p. 115, 2022.

[15] T. Sato, M. Ohzeki, and K. Tanaka, "Assessment of image generation by quantum annealer," *Sci. Rep.*, vol. 11, no. 1, pp. 1–10, 2021.

[16] K. Wadhwani and S. P. Awate, "Controllable image generation with semi-supervised deep learning and deformable-mean-template based geometry-appearance disentanglement," *Pattern Recognit.*, vol. 118, Oct. 2021, Art. no. 108001.
[17] Y. Jiang, S. Yang, H. Qiu, W. Wu, C. C. Loy, and Z. Liu, "Text2human: Text-driven controllable human image generation," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 1–11, 2022.
[18] Y. Wang, K. Jiang, H. Lu, Z. Xu, G. Li, C. Chen, and X. Geng, "Encoder-decoder assisted image generation for person re-identification," *Multimedia Tools Appl.*, vol. 81, no. 7, pp. 10373–10390, 2022.
[19] Z. Ren, X. Y. Stella, and D. Whitney, "Controllable medical image generation via GAN," *J. Perceptual Imag.*, vol. 5, Mar. 2022, Art. no. 000502.
[20] A. Smith and S. Colton, "Clip-guided GAN image generation: An artistic exploration," *Evo**, vol. 2021, p. 17, Mar. 2021.
[21] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on convolutional neural networks (CNN) in vegetation remote sensing," *ISPRS J. Photogramm. Remote Sens.*, vol. 173, pp. 24–49, Mar. 2021.
[22] A. A. Tulbure, A. A. Tulbure, and E. H. Dulf, "A review on modern defect detection models using DCNNs–Deep convolutional neural networks," *J. Adv. Res.*, vol. 35, pp. 33–48, Jan. 2022.
[23] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12116–12128.
[24] S. A. Oprisan, "Interdisciplinary curriculum for computational neuroscience at primarily undergraduate institutions," *J. Comput. Sci.*, vol. 61, May 2022, Art. no. 101642.
[25] E. C. Webber, B. D. McMillen, and D. R. Willis, "Health care disparities and access to video visits before and after the COVID-19 pandemic: Findings from a patient survey in primary care," *Telemed. e-Health*, vol. 28, no. 5, pp. 712–719, 2022.

WEIZHE WANG was born in Wugang, Henan, China, in 1981. He received the master's degree from Zhengzhou University, China. He is currently with the Modern Information Technology College, Henan Polytechnic University. His research interests include video processing and machine learning.

ERZHUANG DAI was born in Zhoukou, Henan, China, in 1995. He received the master's degree from Henan University, China. He is currently with the Modern Information Technology College, Henan Polytechnic University. His research interests include intelligent optimization, big data analysis, and processing.