A Face Recognition Method for Sports Video Based on Feature Fusion and Residual Recurrent Neural Network

Xu Yan
Physical Education Department, North China Electric Power University, Baoding 071003, China
E-mail: yanxu1986ncepu63@163.com

Keywords: face recognition, feature fusion, residual recurrent neural network, reconstruction of ternary confrontation, sports

Received: April 1, 2024

Face recognition technology has penetrated into people's daily life and work fields, and has also been widely applied in sports videos. A video face recognition technology based on feature fusion and residual recurrent neural network is proposed to address the issue of image pose deviation caused by non-cooperative situations. Due to the large number of missing high-frequency data in low resolution facial images, a ternary adversarial reconstruction network is first proposed. It achieves correct image matching through the spatial distance of each image, improving the robustness of the model. For facial recognition in video sequences, higher precision key feature extraction is required. Therefore, this study introduced a residual recurrent neural network to optimize it, and designed its feature fusion and recognition network modules to compensate and extract relevant information before and after frames. Finally, performance verification analysis was conducted on the proposed model, indicating that the recognition accuracy of the recognition system reached 98.3%. In summary, the constructed residual recurrent neural network based on the ternary adversarial reconstruction network framework can effectively achieve video oriented facial recognition.

Povzetek: Predstavljena je metoda za prepoznavanje obrazov v športnih videih, ki temelji na združitvi značilnosti in rekurentnem nevronskem omrežju.

1 Introduction

Face recognition is widely used in fields such as security, payment, and intelligent devices. Traditional portrait recognition for still images and matching facial features is no longer sufficient to meet people's daily needs. More research is focused on non-cooperative dynamic video face recognition scenarios. In real conditions, images often have features such as changes in lighting angle, changes in posture, and masking [1]. In the context of these more challenging circumstances, it is imperative that devices are capable of facial recognition with enhanced speed and precision. The development of modern equipment has a strong demand for face recognition in non-cooperative scenes. For example, high-speed trains or buses in transportation, as well as border controls, require control of visitor access. Face recognition for videos can ensure the security of these places, providing great help for law enforcement personnel. It not only ensures the timeliness of recognition, but also provides early prediction for suspicious behavior. In the commercial field, ordinary people can use facial recognition technology to achieve non-contact payment, improve efficiency, and also provide traffic management for commercial venues. When identifying users, the automatically read customer data can provide a basis for behavioral analysis of the model to achieve personalized recommendations and optimize user experience [2]. In summary, enhancing the recognition performance of realistic image scenes such as posture deviation is the key to the development of dynamic facial recognition technology. This study proposes a Residual Recurrent Neural Network (R-RNN) based on a ternary adversarial reconstruction network framework to solve the problems of face recognition in non-cooperative scenes. The content includes four parts. The first part introduces the current development status of video facial recognition technology. The second part designs and analyzes the framework of the ternary adversarial reconstruction network and the R-RNN. The third part verifies the performance of the recognition model through simulation experiments. The fourth part further summarizes the experimental data.

2 Related works

Facial reconstruction is the key to achieving face recognition in non-cooperative scenes. Deep learning is widely used in the field of visual perception due to its excellent feature extraction performance. Wang et al. believed that the convolutional operators in deep Convolutional Neural Networks (CNN) need to be further improved to overcome the limitations of "local" kernels. Therefore, they proposed a "non-local" model of Speckle Converter (SpT) UNet, with a Pearson correlation

coefficient of 0.989 [3]. Theoharis believed that facial reconstruction is an important part of the branch of computer vision. Compared to 2D data, 3D facial data could better avoid the impact of lighting and pose deviation. A recognition model combining it with CNN was proposed, and the reliability of the model was verified through simulation testing using the Florence dataset [4]. Tewari et al. utilized deep CNN to achieve automatic encoding and achieved 3D facial reconstruction in non-cooperative scenes. The input data of its decoder adopted well-defined code vectors. The encoder extracted useful semantic parameters from a single input image. This facial reconstruction model based on deep CNN had good performance [5]. Dib et al. applied differential ray tracing to facial reconstruction and simulated it under different light intensities using a digital analog light table. Concurrently, the reconstruction optimization equation has been implemented to facilitate reconstruction in complex scenarios, such as self-shadowing, and can also estimate parameters, such as diffuse reflection. Finally, the performance analysis of the model in real scenarios verified its effectiveness [6].

Super resolution and pose correction are two important branches of facial reconstruction. Dastmalchi and Aghaeinia proposed a deep CNN based on pixel loss function for discriminating high-resolution facial images. The Generative Adversarial Network (GAN) was introduced to solve the problem of model over smoothing, and achieved an accuracy of 86.1% in the LFW dataset [7]. Nagar et al. proposed the use of position blocks for facial super-resolution optimization to address the impact of Gaussian pulse noise on low resolution images. This was because ordinary facial super-resolution methods are highly susceptible to noise. Its principal component analysis could analyze the matrix of pixel noise details and eliminate pulse noise. Residual learning was used to update the training set and weaken Gaussian noise, and the effectiveness of this method has been verified [8]. Teng et al. proposed alternating improvement algorithms to address the issue of insufficient accuracy in deep learning facial reconstruction, especially for low resolution images. This algorithm improved network performance through alternating training of dual convolutional networks, which are used for facial reconstruction and attribute correction, respectively. Finally, the reliability of this method was demonstrated in the CelebA dataset [9]. Sharma used GAN for facial recognition to enhance the performance of super-resolution images, and experiments has shown that its error rate was only 0.001% [10].

In conclusion, deep learning can be employed for video facial recognition. However, the existing state-of-the-art (SOTA) methods in the literature still exhibit deficiencies in their ability to cope with the aforementioned complex situations, particularly in terms of recognition performance in situations involving posture deviation and non-cooperative scenarios. Meanwhile, SOTA models often perform well under laboratory conditions, but in the real world, their robustness is insufficient due to the variability of poses and the unpredictability of non-cooperative scenarios. The paper selects the ternary GAN as the fundamental framework for facial reconstruction, which directly provides innovative solutions to the challenges in this field. This approach facilitates the advancement of facial recognition technology in non-cooperative scenarios, particularly in enhancing recognition accuracy, optimizing model generalizability, and accelerating real-time processing capabilities. The study also utilizes an R-RNN based on feature fusion as a facial recognition model. The utilization of R-RNNs to optimize the feature fusion process, which combines the advantages of triplet loss and Recurrent Neural Networks (RNN), is employed to enhance the robustness of the model to occlusion and lighting changes. Table 1 shows the summary of the related works.

Table 1: Summary of the related work

Researchers	Key contributions	Models or techniques used	Main results or performance indicators
Wang et al. [3]	Propose a "non-local" model for SpT UNet	Deep CNN	Achieved a Pearson correlation coefficient of 0.989
Theoharis [4]	Propose a 3D facial data recognition model combined with CNN	3D facial data and CNN	Model reliability validated through simulation tests
Tewari et al. [5]	Realizing 3D facial reconstruction in non-cooperative scenarios	Automatic encoding of deep CNN	Model demonstrated good expressiveness
Dib et al. [6]	Applying Differentiable Ray Tracing to Facial Reconstruction	Differentiable Rendering	Achieved reconstruction under complex conditions like self-shadowing
Dastmalchi and Aghaeinia [7]	Propose a deep CNN based on pixel loss	Deep CNN + GAN	Achieved an accuracy of 86.1% in the LFW

	function		dataset
Nagar et al. [8]	Propose using position blocks for facial super-resolution optimization	Principal Component Analysis (PCA) + Residual Learning	Mitigated the impact of Gaussian impulse noise on low-resolution images
Teng et al. [9]	Propose alternating improvement algorithms to enhance the accuracy of facial reconstruction	Dual Convolutional Networks with alternating training	Method's reliability validated in the CelebA dataset
Sharma [10]	Using GAN for Facial Recognition	GAN	Experiments showed an error rate of only 0.001%
The research of this article	Propose a Face Residual Recurrent Neural Network (FR-RNN) model based on the TL-GAN framework to optimize face recognition in non-cooperative scenarios	TL-GAN+FR-RNN+Tensorflow	The TL-GAN+FR-RNN model has an accuracy of up to 98.3% in facial recognition tasks and performs best on the IJB-A dataset, with an accuracy of 96.3%

3 A face recognition method for sports video based on feature fusion and R-RNN

The basic framework of a recognition network model based on ternary adversarial reconstruction is studied, aiming to solve the problem of face recognition in non-cooperative states. This model uses a GAN as the basic architecture, and introduces a ternary adversarial reconstruction recognition network for construction. Finally, the GAN is used for training. The proposed optimization construction of video facial recognition technology based on feature fusion R-RNN includes two modules: feature fusion and facial recognition. This study also utilizes R-RNN to improve the accuracy of feature fusion [11, 12].

3.1 Construction of a basic framework for reconstructing and identifying network models based on ternary confrontation

Facial recognition technology based on video clips is widely used. In the field of sports, this technology can be used to achieve facial recognition and violation detection functions. However, facial recognition functions in non-cooperative states often experience a sudden decrease in recognition accuracy due to issues such as posture deviation and clarity. Therefore, the design of a facial reconstruction recognition system that addresses the aforementioned defects is very necessary and has research prospects. This study is based on the GAN architecture and introduces the concept of feature mapping to achieve multi-pose facial correction. This type of network can reduce its dependence on supervised learning and also improve computational accuracy. The basic structure is shown in Figure 1.

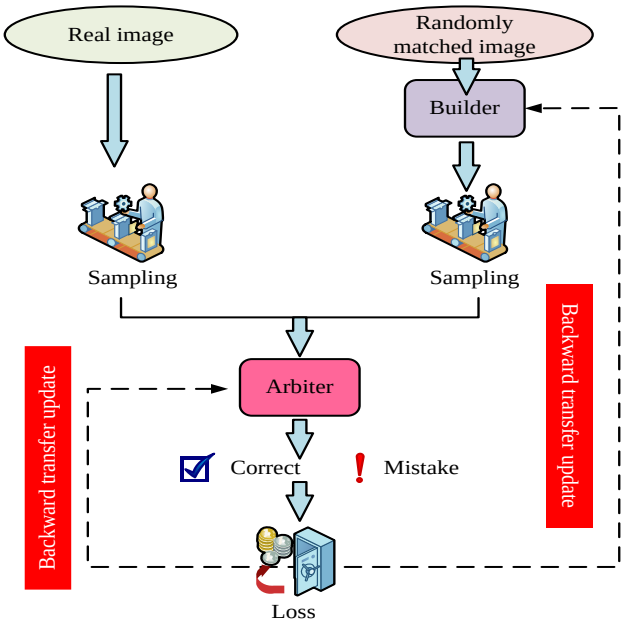


Figure 1: The basic structure of GAN

The basic principle of GAN is the mutual confrontation between the generator and discriminator for training. When facing low resolution side face images, the lack of high-frequency data often leads to feature similarity issues after facial correction. Therefore, the introduction of the triplet loss theory can construct a Triplet Loss Constrained GAN for Reconstruction and Recognition (TL-GAN). The principle of distance measurement is shown in Figure 2.

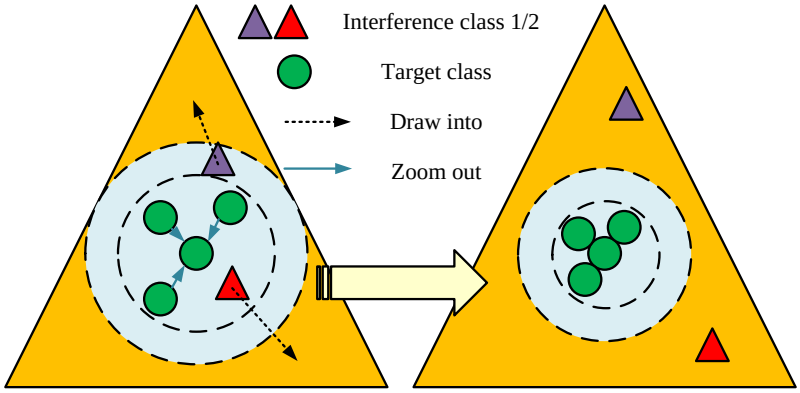


Figure 2: Principle of distance measurement in image recognition

The basic principle of the model is to achieve accurate matching of the same face through spatial distance in high-dimensional space. It is divided into three major modules, namely low-resolution correction module, super-resolution module, and discrimination module. The first two are combined to generate a network. Low-resolution pose correction uses convolutional and deconvolution networks. The convolutional layer and adaptive attention module respectively achieve feature extraction and detail data acquisition. The input-output relationship of the reconstructed network of the codec combination is shown in equation (1).

$$I^{HR} = F_{dec}(F_{enc}(I^{LR})) \tag{1}$$

In equation (1), I^{LR} represents the network input value. I^{HR} represents the reconstructed network output value. F_{dec} and F_{enc} are decoder and encoder, respectively. The input output relationship of the recognition network is shown in equation (2).

$$\text{Identity discrimination} = F_{class}(I^{HR}) \tag{2}$$

In equation (2), F_{class} represents the classification recognition network. Ordinary triplet losses have uncertainty. The difference in positive and negative examples can cause defects such as under-fitting in the model. The shortest distance ternary loss function L_{triple} is introduced for optimization, as shown in equation (3).

$$L_{triple}=\sum_{i=1}^{N}\left(\begin{array}{c}\left\|F_{enc}(I_i^{LR})-F_{enc}(I_i^{LR+})\right\|^2-\left\|F_{enc}(I_i^{LR})-F_{enc}(I_i^{LR-})\right\|^2\\+\frac{1}{2}\left\|F_{enc}(I_i^{LR})-F_{enc}(I_i^{LR-})\right\|^2\end{array}\right) \quad (3)$$

In equation (3), $i=1,2,...,N$ represents the serial number of the portrait. The corresponding triplet symbol is represented as $(I_i^{LR}, I_i^{LR+}, I_i^{LR-})$, which is the low-resolution profile image and low-resolution frontal image of the portrait, as well as the low-resolution frontal image of another person. The vector features of an image are mapped through an encoder. The function uses the distance between vectors for similarity recognition. The construction of a triplet is a random pattern, which may cause the divergence of the negative target I_i^{LR-} to be too high and ultimately slow down the convergence speed of the model. The method of selecting I_i^{LR-} in this study is shown in equation (4).

$$I_i^{LR-}=\arg\min\left\|F_{enc}(I_i^{LR+})-F_{enc}(I_i^{LR})\right\|^2 \quad (4)$$

After selecting faces with a focus on distinguishing similar features, the training speed and fitting degree of the model can be improved to a certain extent. The training process of the model is shown in Figure 3.

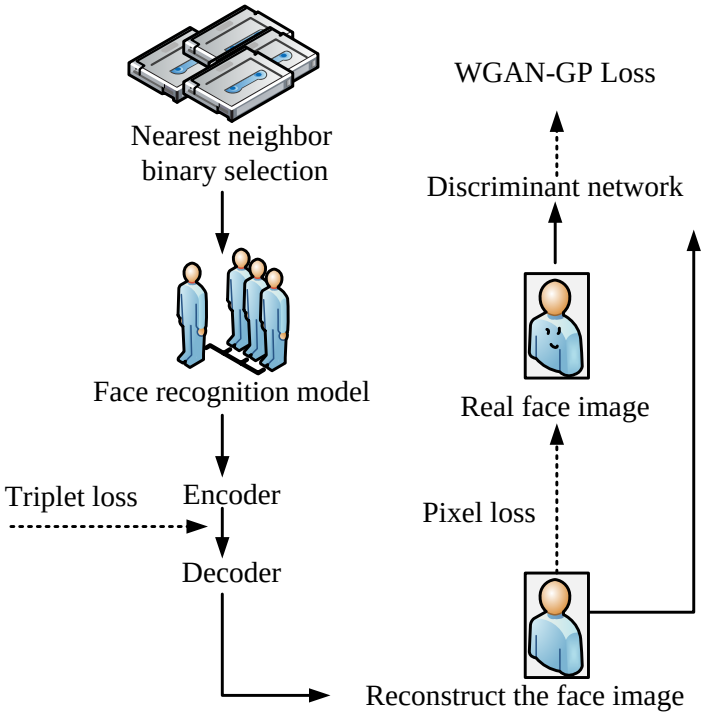


Figure 3: Training flow of ternary adversarial network

The triplet image enters the feature space through an encoding network. The model utilizes the shortest distance triplet loss to constrain its spatial distance, and also introduces pixel loss to further enhance the model's perception of the front face. GAN is the final training platform. The discriminator structure is WGAN-GP. In summary, the target loss of the model is the sum of multiple losses, as shown in equation (5).

$$L^{SR}=\alpha L_{pixel}+\gamma L_{triple}+\beta L_{WGAN-GP} \quad (5)$$

In equation (5), L_{pixel} represents pixel loss. $L_{WGAN-GP}$ represents WGAN-GP loss. α, γ and β respectively represent the weights of the corresponding losses. Usually, the weight value of the latter two losses is higher. After continuous training, the objective function of ordinary GAN will cause the gradient to disappear. The bulldozing distance can be employed to enable the feature vector input after feature extraction to counteract losses, thereby facilitating continuous optimization of the generator and discriminator. This, in turn, stabilizes the network structure. When the discriminator compares images, the WGAN-GP loss will use the distance between each data to determine

similarity, as shown in equation (6).

$$L_{WGAN-GP} = D(I^{HR}) - D(I^{HR+}) + \lambda(\left\|\nabla D(I^{HR})\right\|_2 - 1)^2 \quad (6)$$

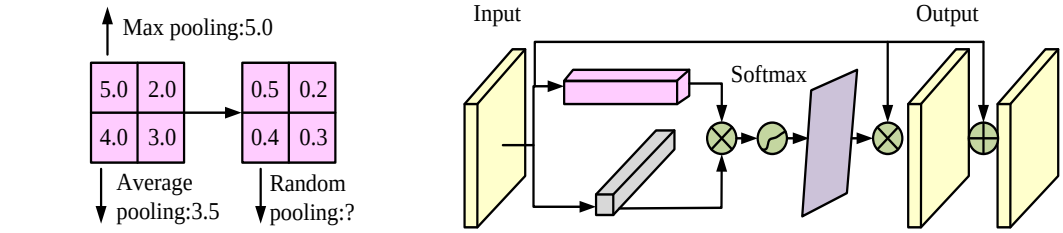
In equation (6), D represents the output data of the discriminator. I^{HR} represents the reconstruction of the front face image. I^{HR+} represents a true facial image. λ is a pre-set value for the user. Pixel loss can be used to constrain surface similarity, as shown in equation (7).

$$L_{pixel} = \frac{1}{3m^2}\sum_{r=1}^{3}\sum_{i=1}^{m}\sum_{j=1}^{m}\left(I_{i,j,r}^{HR+} - I_{i,j,r}^{HR}\right)^2 \quad (7)$$

In equation (7), λ is the number of samples in the training set. $i/j/r$ are different categories.

3.2 Optimization and construction of video facial recognition technology based on feature fusion R-RNN

Although the above framework can achieve optimized facial reconstruction, the extraction of key features in videos is not precise enough. Therefore, the FR-RNN is introduced to optimize it. The model can be roughly divided into two modules: feature fusion and face recognition. The difference between video super-resolution and ordinary image super-resolution lies in the strong correlation between the front and back frames of the former, so feature fusion is necessary. Feature fusion essentially involves supplementing information from images that are interrelated, so as to enhance their data expression capabilities. The feature fusion technology based on deep learning is superior to traditional fusion technologies, including the fusion between feature maps [13]. The pooling layer is one of the manifestations of feature map fusion. The self attention mechanism belongs to one of the manifestations of fusion between feature maps, as shown in Figure 4.



(a)Pooling layer feature fusion framework (b)Basic framework of attention mechanism model

Figure 4: Basic framework of pooling layer and attention mechanism feature fusion

In Figure 4 (a), when the image passes through the pooling layer, the average pooling will select the average feature value. The maximum value is selected for the maximum pooling layer. Random pooling involves selecting any value through a probability matrix. Figure 4 (b) actually shows the connection implemented in the channel. The attention mechanism achieves important data extraction by training the adaptive weight matrix. However, this method only targets the connection of a single image and is not suitable for feature map fusion in videos. Common feature fusion techniques for video image super-resolution include 2/3D convolution and RNN. The difference between 2/3D convolutional fusion is mainly reflected in the dimension of feature fusion. The former is directly connected and fused at the channel. The latter takes video as input and connects it in both spatial and temporal dimensions. RNN needs to calculate the context correlation of video images, and the resulting hidden states can be connected to the current frame [14, 15]. The feature description performance of ordinary images in the network has been significantly improved. The most important thing at present is to optimize the feature fusion of video frame images to enable them to extract facial features more accurately. To enhance the model's ability to perform feature fusion, this study analyzes and compares the three fusion technologies, as shown in Figure 5.

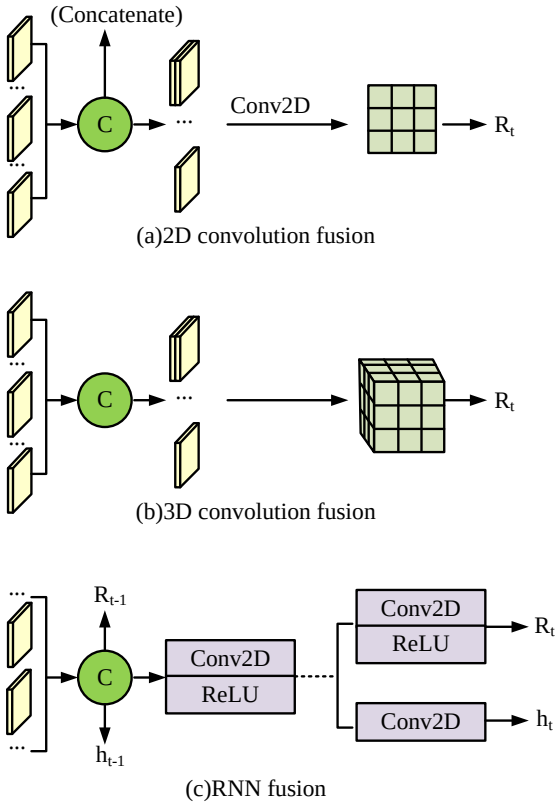


Figure 5: Different fusion technology frameworks

In Figure 5, 2D convolutional fusion is shown in equation (8).

$$R_t = W_{net2D}\left\{Concatenate\left[I_{t-T},...,I_{t+T}\right]\right\} \quad (8)$$

In equation (8), $Concatenate\left[I_{t-T},...,I_{t+T}\right]$ represents the cascading operation of the sequence image. The dimensional data is represented as $NC \times H \times W$. $N = 2T + 1$ represents the length of the video sequence, which is the number of input image channels. λ represents the length and width dimensions of the image. The 3D convolutional fusion is shown in equation (9).

$$R_t = W_{net3D}\left\{Concatenate\left[I_{t-T},...,I_{t+T}\right]\right\} \quad (9)$$

In equation (9), the convolutional kernel becomes three-dimensional, and its motion on the spatio-temporal axis is achieved by inputting a video sequence, facilitating its extraction of spatio-temporal feature data. RNN utilizes 2D convolutional encoding to achieve fusion and obtain the output of the current frame and the hidden state of the subsequent frames. The fusion technology of the first two can better achieve feature fusion when faced with a small number of sequences, but the increase in sequence length will ultimately lead to computational difficulties. On the contrary, the input of RNN is only the pre and post frame data, and the fusion is achieved by utilizing the hidden states of the two. This recursive method is more suitable for recognizing video images with longer sequences. Therefore, using this technology for feature fusion is the most suitable. FR-RNN can further improve the potential gradient vanishing defects in feature fusion. Its basic structure is shown in Figure 6.

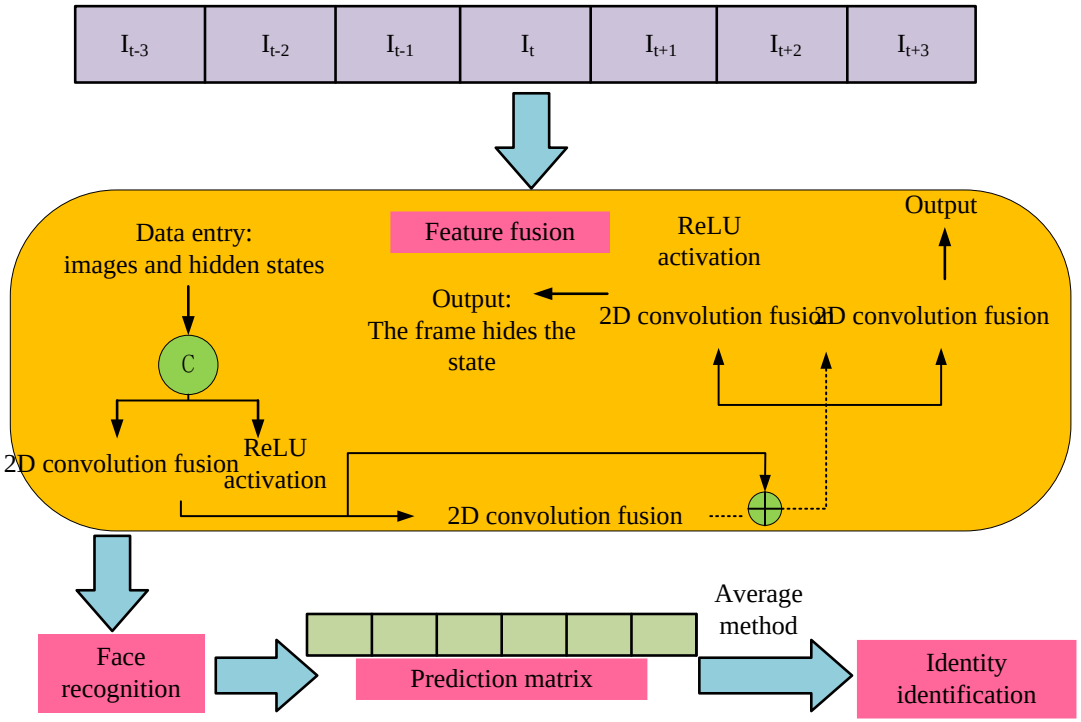


Figure 6: Overall input value of FR-RNN

In Figure 6, the overall input value of FR-RNN is a video image sequence, with dimensions of $m' \times m'$. After feature fusion, the recognition network can obtain the discriminative data of the current frame, as shown in equation (10).

$$R_t = W_{conv2D}\{\hat{x}_k\} \tag{10}$$

In equation (10), $\hat{x}_k$ represents the output value of the current frame. $\hat{x}_k$ represents the encoding of the channel connection, and its specific calculation is shown in equation (11).

$$\begin{cases} \hat{x}_k = \hat{x}_{k-1} + \Im(\hat{x}_{k-1}), k \in [1, K] \\ \hat{x}_0 = \sigma\left(W_{con2D}\{[I_{t-1}, I_t, o_{t-1}, h_{t-1}]\}\right) \\ h_t = \sigma\left(W_{con2D}\{\hat{x}_k\}\right) \end{cases} \tag{11}$$

In equation (11), $\hat{x}_k$ represents the channel connection encoding for the next frame. K represents the number of standard residual blocks. $\hat{x}_0$ represents the connection of the four parameters on the channel, namely the hidden state of the previous frame, the output of the previous frame, the input of the current frame, and the input of the previous frame. h_t represents the hidden state of the current frame. $\Im(\hat{x}_{k-1})$ represents the final residual block output. The prediction matrix obtained through feature fusion and recognition is shown in equation (12).

$$IP_t = F_{class}(R_t) \tag{12}$$

In equation (12), IP_t represents the final prediction matrix. F_{class} represents the recognition network. Among them, Light Convolutional Neural Networks (LightCNN) and Visual Geometry Group Face (VGG-Face) are two common facial recognition network models. Subsequently, the averaging method is used to process the prediction matrix obtained in the previous text, and the final output of the recognition model can be obtained, as shown in equation (13).

$$\text{Identity discrimination} = \frac{1}{N}\sum_{t=1}^{N} IP_t \tag{13}$$

The training of FR-RNN includes feature fusion for each frame and training for recognition modules. By introducing cross entropy to construct a loss function and calculating the deviation between the predicted label vector and its true value, equation (14) can be obtained.

$$Loss = \frac{1}{n}\sum_{x}\left[y \log \hat{y} + (1-y)\log(1-\hat{y})\right] \tag{14}$$

In equation (14), x represents the sample. n represents the number of samples. $\hat{y}$ represents the predicted label vector. y represents the actual label vector. Video facial recognition is a classification problem. Softmax regression is introduced to process it, as shown in equation (15).

$$L_{soft\max} = \frac{1}{m_1}\sum_{i=1}^{m_1} \log \frac{e^{W_{y_i}^T x_i + by_i}}{\sum_{j=1}^{n_1} e^{W_j^T x_i + bj}} \tag{15}$$

In equation (15), m_1 represents the batch size. n_1 represents the number of classes. y_i represents the specific category. x_i represents the i depth features under the corresponding category. W_j represents the j column of the weight. b represents deviation. $e^{W_j^T x_i + bj}$ represents fully connected layer output. The increase in

the output value weight of the fully connected layer can reduce model losses.

Overall, the architecture of the FR-RNN model is based on generating GANs, which includes generators and discriminators for generating and discriminating images, and is trained through the adversarial process between them. To enhance the model's ability for facial image reconstruction, especially when processing low-resolution images, TL-GAN is introduced in the study, which utilizes the triplet loss theory to improve the model performance. The FR-RNN model further integrates R-RNN, which consist of two modules: feature fusion and face recognition. The main responsibility is to optimize the accuracy of feature fusion to better process keyframe features in video sequences. The standard practice based on GAN adopts convolutional layers, pooling layers, and ReLU activation functions. In terms of selecting the loss function, the FR-RNN model combines pixel loss, WGAN-GP loss, and triplet loss, among which WGAN-GP loss is particularly used to improve the stability of the training process and avoid the problem of gradient vanishing during the optimization process.

4 Performance simulation analysis of feature fusion and FR-RNN model in video face recognition performance

The performance simulation experiment of the video facial recognition model is divided into two parts, which are the analysis of each module and the overall analysis. The performance analysis of the model itself includes four aspects: Structural Similarity (SSIM), recognition accuracy, rank N, and model size. The comparative analysis of the models conducted experiments on the recognition accuracy of each model [16-18].

4.1 Performance verification analysis of TL-GAN framework and FR-RNN module

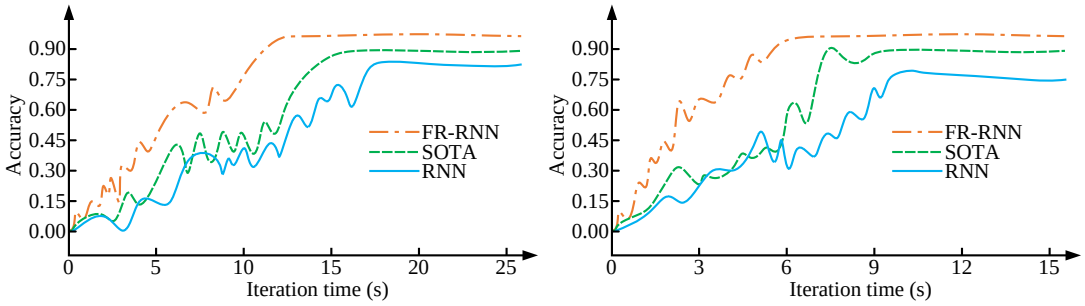
Table 2 shows the parameters of the experimental environmen.

Table 2: Experimental environment and parameter settings

Experimental environment	Parameter setting
Graphics card	GTX1080Ti
Operating system	Linux
Deep learning framework	Tensor flow
Pre-treatment method	Double - and three-wire interpolation
Image size	32×32
$\alpha / \gamma / \beta$	0.01/0.1/0.1
Learning rate	0.001
Weight attenuation	0.01
Batch	20
Stochastic gradient optimization	Adam (β_1 =0.9, β_2 =0.999)
Preconditioning	Double trilinear interpolation method
Data Sets	Multi-PIE / IJB-A

The study uses data augmentation techniques such as rotation, scaling, cropping, and color transformation. The batch size used during the training process is 20, and the Adam optimizer is used with parameters B1=0.9 and B2=0.999. Then, performance analysis is conducted on the image-based ternary adversarial reconstruction recognition network. The 250 participants in Session01 are selected from 6 different angles under the same lighting and facial expressions, and allocated in a 4:1 ratio as the training and testing sets, respectively. The data preprocessing steps include using the double trilinear interpolation method to process images to improve image quality and prepare for subsequent facial recognition analysis. All images are uniformly adjusted to a size of 32 ×32 pixels, and this standardization process helps to accelerate model training speed and reduce computational resource consumption. A comprehensive data distribution strategy has been devised for the multi-PIE dataset with the objective of ensuring the diversity of images in terms of angles and lighting conditions. This approach aims to simulate the various challenges that face recognition may encounter in the real world. The paper compares the

accuracy of the proposed FR-RNN algorithm with SOTA and traditional RNN algorithms, as shown in Figure 7.



(a) The first group of experiments (b) Second set of experiments

Figure 7: Comparison of accuracy of different algorithms

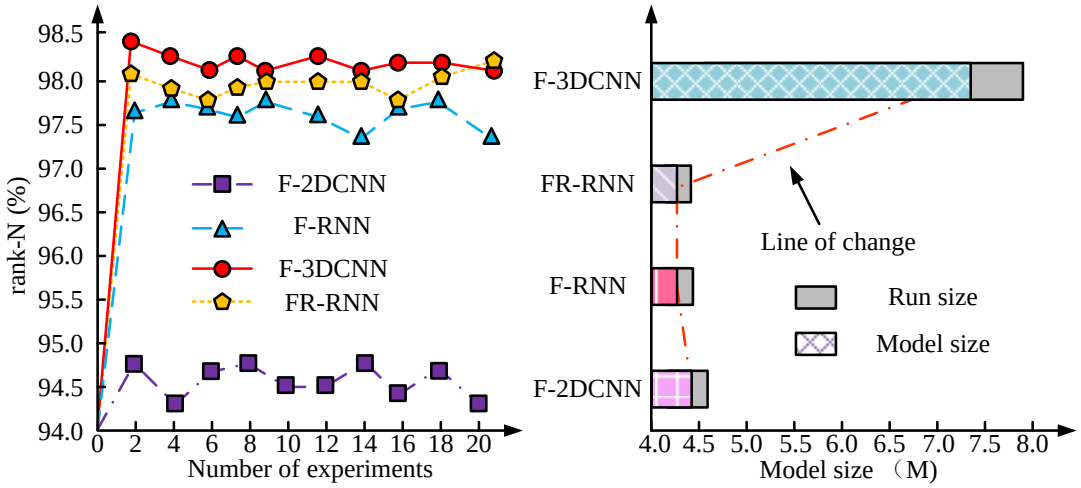
As shown in Figure 7 (a), in the first set of tests, the accuracy of the proposed FR-RNN algorithm reaches 94.3, while the accuracy of the SOTA algorithm is 89.7, and the accuracy of the traditional RNN algorithm is 82.5. In Figure 7 (b), in the second set of tests, the accuracy of the proposed FR-RNN, SOTA and traditional RNN algorithms reaches 94.5, 89.4 and 81.2. This indicates that the FR-RNN algorithm is more effective in handling video facial recognition tasks. This study introduces commonly used Two Path Generative Adversarial Network (TP-GAN), Factorization Machines Deep Neural Network (FNM), LightCNN, and VGG-Face as controls. SSIM and recognition accuracy are used as evaluation indicators for model performance. Table 3 shows the experimental results.

Table 3: Performance comparison of portrait reconstruction model

Index	Moldel	Angle /°					
		±15	±30	±45	±60	±75	±90
SSIM	TL-GAN	0.7105	0.6643	0.6512	0.6327	0.6249	0.6078
	TP-GAN	0.6987	0.6541	0.6276	0.6048	0.5809	0.5699
	FNM	0.6847	0.6362	0.6009	0.5806	0.5462	0.4621
Accuracy rate /%	LightCNN	87.76	85.73	69.42	30.64	10.34	2.13
	VGG-Face	89.81	87.76	71.45	32.74	12.42	4.15
	TL-GAN+LightCNN	98.16	95.92	93.88	91.84	85.72	71.44
	TL-GAN+VGG-Face	98.16	96.95	94.91	93.86	87.73	74.77
	TP-GAN+LightCNN	88.71	88.08	85.42	77.73	67.45	54.68
	FNM+LightCNN	94.62	92.51	89.77	85.31	77.25	61.21

Due to the large amount of data, the study only conducts SSIM performance comparison analysis on the TP-GAN, FNM, and TL-GAN models. As the angle decreases, the facial reconstruction ability of each model will also be correspondingly improved. FNM has the lowest SSIM. The average SSIM value of TL-GAN is 0.6486, which is 9.79% higher than the average SSIM value of FNM. At ±90°, the SSIM of TL-GAN is 6.23% and 23.97% higher compared to TP-GAN and FNM, respectively. The facial reconstruction image of

TP-GAN is relatively clear, but there may be artifacts in the image. The FNM facial reconstruction image has relatively more detailed features, but it is obvious that as the angle increases, its facial correction performance will decrease, so the model has higher requirements for lighting. TL-GAN can maintain relatively stable facial correction performance, with SSIM values only differing by 0.1027 under $\pm 90^\circ$ and $\pm 15^\circ$ conditions. Therefore, this model can better extract correct detail features and achieve more accurate facial correction. In the accuracy verification experiment of each model, LightCNN, VGG-Face are combined with the other three models. Experiments have shown that the recognition performance of LightCNN and VGG-Face alone is much lower than that of other models. Especially at $\pm 90^\circ$, the recognition accuracy is lower than 5%. The average recognition accuracy is 47.59% and 49.64%, respectively. The average recognition accuracy of TL GAN+LightCNN, TL GAN+VGG-Face, TP GAN+LightCNN, and FNM+LightCNN are 89.49%, 91.06%, 76.89%, and 83.41%, respectively. Therefore, the combination of TL-GAN and other models has the best performance, reaching over 89%. At $\pm 90^\circ$, the recognition accuracy of TL-GAN+VGG-Face is 1.57%, 14.17%, and 7.65% higher than that of TL-GAN+LightCNN, TP-GAN+LightCNN, and FNM+LightCNN models, respectively. This study selects the IJB-A, YTC, and YTF datasets as experimental samples to validate the performance of the FR-RNN model. There is high-quality frontal data and corresponding video sequences in IJB-A, which can be used to simulate sports videos. Due to its low-resolution and multi-pose data features that are very similar to actual video surveillance, it is a better choice for recognition verification. The YTC and YTF datasets lack high-quality frontal images. Therefore, this study utilizes the FaceChoose algorithm for high-quality image selection and skipping. This study first fixes the number of residual blocks K to 5 and conducts experiments on each model separately. The experimental results are shown in Figure 8.

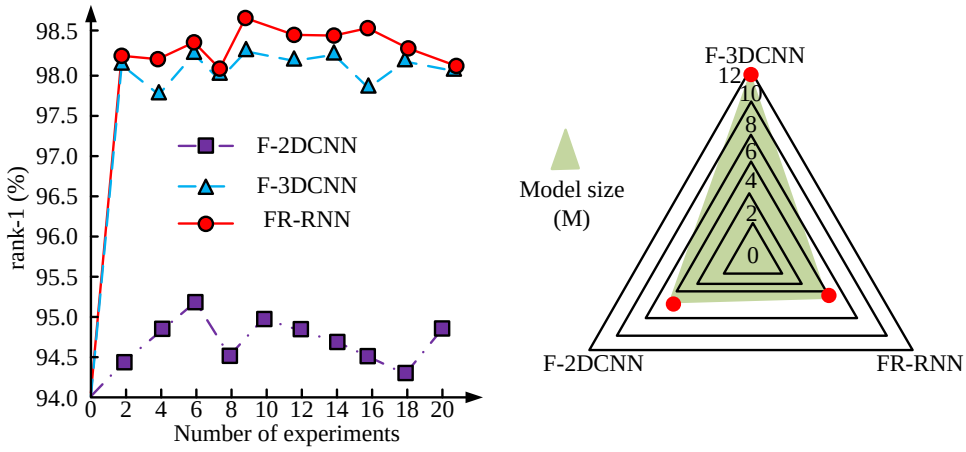


(a) The recognition rate performance of each model

(b) Size of each model

Figure 8: Comparison of performance of each recognition model when K=5

The above models all use the same training dataset and test set sequence length. Feature extraction is unified as a residual block module. The data ratio for training and testing is set to 8:2. Figure 7 compares the rank N recognition rate and model size of each model. The rank N recognition rate represents the proportion of correct attributes among the top N model recognition results. The size of the model indirectly reflects the parameter quantity and computational speed of the model. In Figure 8 (a), under the condition of K=5, F-3DCNN has the highest recognition accuracy, with an average of 98.2%, which is 3.6% and 2.7% higher than F-2DCNN and F-RNN, respectively. This indicates that the accuracy of the model is relatively excellent. The average rank N value of FR-RNN is 98.0%, which is only 0.2% lower than F-3DCNN. Therefore, the difference in recognition accuracy between the two is not significant. In Figure 8 (b), the average size of F-3DCNN is 11.7M, while the average size of F-2DCNN, F-RNN, and FR-RNN models is 4.4, 4.2 and 4.2, respectively. This indicates that although F-3DCNN has the highest recognition accuracy, its operating speed is much lower than other models. Based on the two-indicator data, FR-RNN has good comprehensive recognition performance. The experiment reset the K value to 10, and the experimental results are shown in Figure 9.



(a) The recognition rate performance of each model

(b) Size of each model

Figure 9: Comparison of performance of each recognition model when K=5

In Figure 9, there is no experimental data for F-RNN, as the model experienced gradient vanishing during training. This also indirectly confirms the effectiveness of residual connections in improving model performance. In Figure 9 (a), as the number of residual blocks increases, the average recognition rate of FR-RNN is higher than that of F-3DCNN. Its average rank N value is 98.5%, which is 3.7% and 0.4% higher than the F-2DCNN and F-3DCNN models, respectively. In Figure 9 (b), the size of FR-RNN has increased to some extent, but it is also at its minimum value. The three model sizes are 5.7M, 5.9M, and 11.6M, respectively. This is because the F-2DCNN and F-3DCNN models utilize the increase in model parameter quantity to achieve the processing of long sequence data. FR-RNN uses recursive methods for feature fusion processing. The correlation between model size and sequence length is weak, which makes it easier to handle real-time and video sequence data, while avoiding gradient vanishing in hidden states, ensuring the stability of sequence length.

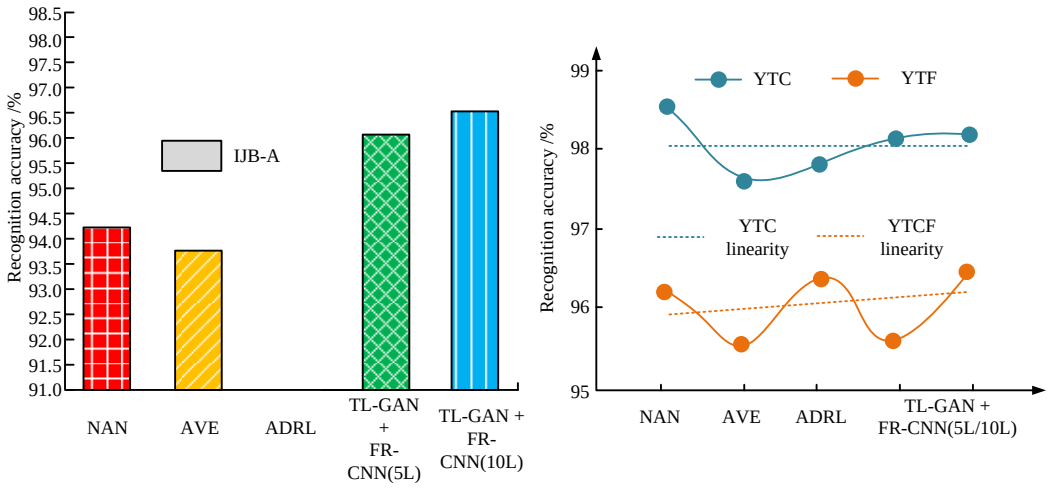
4.2 Performance verification and comparative analysis of recognition models based on TL-GAN framework and FR-RNN module

This study further analyzes the stability of the overall model. By setting different sequence lengths and residual blocks, it is determined whether the differences in the model are too large, as a way to determine the stability of the model. Table 4 shows the experimental results.

Table 4: Verifies the stability of the overall model

	Parameter	Recognition accuracy
Frame number	5	95.3%
	10	96.9%
	15	97.1%
	20	97.5%
	25	97.9%
Number of residual blocks	3	92.5%
	4	92.9%
	5	94.6%
	6	94.7%
	7	95.2%
	8	95.3%
	9	95.1%
	10	95.0%

In Table 4 above, the proposed TL-GAN+FR-RNN model is better at processing long sequence data, as shown in the data of frame number and recognition accuracy. As the number of frames increases, its accuracy also improves. The difference in recognition accuracy between models with frame numbers 5 and 25 is 2.6%. However, when the number of frames increases to a certain limit, the amplitude of accuracy increase will decrease. The model recognition accuracy difference between frame 20 and frame 25 is only 0.4%. Experiments have shown that TL-GAN+FR-RNN can also achieve high-precision recognition and better feature fusion when facing changes in data frame numbers. According to the data in Table 4 on the number of residual blocks and the accuracy of model recognition, it can be concluded that an appropriate increase in the number of residual blocks can have a positive impact on the performance of model recognition. There is also a phenomenon of maintaining stability after increasing the limit value. The recognition accuracy for residual blocks of 3 and 8 differs by 2.8%. When the number of residual blocks is between 7 and 10, its recognition accuracy remains stable in the [95, 93] range. Based on 95%, the average deviation is 0.15%. Therefore, the change in the number of residual blocks does not affect the stability of the overall model. Accumulated residual blocks can actually improve the accuracy of the model to a certain extent. Neural Aggregation Network (NAN), Attention Deep Reinforcement Learning (ADRL), and Template Depth Reconstruction Model (TDRM) are introduced and compared with the research algorithm, as shown in Figure 10.



(a) Model performance in IJB-A dataset (b) Model performance in YTC/YTF dataset

Figure 10: Comparison of recognition accuracy of various face recognition models

In Figure 10, in the IJB-A dataset, the TL-GAN+FR-RNN model with 10 residual blocks has the highest accuracy of 96.3%. According to the order in the figure, it is 2.1%, 2.7%, and 0.5% higher than the other models, respectively. The phenomenon of gradient disappearance occurred in ADRL. In YTC data, the recognition rates of each model do not differ significantly, with a mean of 98%. The research method is slightly lower than the NAN model by 0.3%. In the YTF model, the difference in recognition rates among different models is still small. The accuracy of the TL-GAN+FR-RNN model reaches 96.2%, which is 1% higher than the TDRM model. In summary, the TL-GAN+FR-RNN model can output more features for hidden states, achieving optimization of recognition rate.

5 Discussion

The FR-RNN model based on the TL-GAN framework proposed in the study has demonstrated superior performance in video facial recognition tasks, especially in non-cooperative scenarios. Simulation analysis shows that the recognition accuracy of the model on the IJB-A dataset reaches 96.3%, which is outstanding in current research and surpasses the SOTA methods in existing literature. For example, although the SpT UNet model proposed by Wang et al. [3] achieved a Pearson correlation coefficient of 0.989, it was not as accurate in facial recognition as the proposed model. Although Theoharis T's 3D facial recognition model has been validated for reliability through simulation testing, there are limitations in processing video sequence data. Although the deep CNN proposed by Dastmalchi and Aghaeinia [7] achieved an accuracy of 86.1% on the LFW dataset, the model demonstrated higher performance in more complex video face recognition tasks.

The performance differences may be mainly attributed to several key factors. Firstly, the TL-GAN framework effectively combines the advantages of triplet loss and GAN to better handle attitude deviation and lighting changes. Secondly, the FR-RNN model optimizes feature fusion and enhances feature expression ability through residual loop mechanism. Finally, the training strategy

adopted, including pixel loss and WGAN-GP loss, helps to improve the robustness and accuracy of the model.
The proposed model provides novel contributions to the field of facial recognition, particularly in optimizing facial recognition in non-cooperative scenarios, processing long sequence data, and considering real-time performance and computational efficiency. These characteristics render the model not only innovative in theory but also potentially valuable in practical applications, particularly in face recognition tasks that necessitate the processing of complex scenes and long sequence data.

6 Conclusion

To further improve facial recognition technology for videos, this study proposed a FR-RNN model based on the TL-GAN framework. The purpose was to solve the problem of low-recognition accuracy caused by attitude deviation and lighting in non-matching images. The simulation analysis of the model showed that in the experiment of the TL-GAN framework, the average SSIM was 0.6486, which was 9.79% higher than the FNM model. In the experiment on the FR-RNN model, when K=5, the rank N mean of the FR-RNN model was 98.0%, which was only 0.3% lower than the highest recognition rate of the F-3DCNN model. Its model size was lower than 7.5M, so its overall performance was the best. In the verification of the overall model stability, when the number of residual blocks was between 7-10, its recognition accuracy remained stable in the [95,93] range. Based on 95%, the average deviation was 0.15%. In the comparison between TL-GAN+FR-RNN and other models, in the IJB-A dataset, the accuracy of TL-GAN+FR-RNN using 10 residual blocks was 96.3%, which was 2.7% higher than the ADRL model. This had always been at a high level in other datasets, with the best overall performance. However, there are still some shortcomings in the experiment, such as reducing model complexity and improving computational speed. At the same time, the experiment also needs to further apply the model to capture multiple faces to adapt to actual scene requirements.

References

[1] D. Tang, and J. Hao, “A deep map transfer learning method for face recognition in an unrestricted smart city environment,” Sustainable Energy Technologies and Assessments, vol. 52, no. 8, pp. 102207-102215, 2020. https://doi.org/https://10.1016/j.seta.2022.102207

[2] F. Zhang, N. Liu, L. Chang, F. Duan, and X. Deng, “Edge-guided single facial depth map super-resolution using CNN,” IET Image Processing, vol. 14, no. 17, pp. 4708-4716, 2021. https://doi.org/https://10.1049/iet-ipr.2019.1623

[3] Y. Wang, H. Wang, and M. Gu, “High performance "non-local" generic face reconstruction model using the lightweight Speckle-Transformer (SpT) UNet,” Advances in Optoelectronics, vol. 6, no. 2, pp. 220049-220058, 2023. https://doi.org/10.29026/oea.2023.220049

[4] T. Theoharis, “Robust 3D face reconstruction using one/two facial images,” Journal of Imaging, vol. 7, no. 9, pp. 169-176, 2021. https://doi.org/https://10.3390/jimaging7090169

[5] A. Tewari, M. Zollhofer, F. Bernard, P. Garrido, H. Kim, P. Perez, and C. Theobalt, “High-fidelity monocular face reconstruction based on an unsupervised model-based face autoencoder,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 357-370, 2020. https://doi.org/https://DOI:10.1109/TPAMI.2018.2876842

[6] A. Dib, G. Bharaj, J. Ahn, C. Thébault, P. H. Gosselin, M. Romeo, and L. Chevallier, “Practical face reconstruction via differentiable ray tracing,” Computer Graphics Forum, vol. 40, no. 2, pp. 153-164, 2021. https://doi.org/https://DOI:10.1111/cgf.142622

[7] H. Dastmalchi, and H. Aghaeinia, “Super-resolution of very low-resolution face images with a wavelet integrated, identity preserving, adversarial network,” Signal Processing. Image Communication: A Publication of the European Association for Signal Processing, vol. 1, no. 107, pp. 116755-116767, 2022. https://doi.org/https://DOI:10.1016/j.image.2022.116755

[8] S. Nagar, A. Jain, P. K. Singh, and B. AK, “Mixed-noise robust face super-resolution through residual-learning based error suppressed nearest neighbor representation,” Information Sciences, vol. 1, no. 546, pp. 121-145, 2021. https://doi.org/https://DOI:10.1016/j.ins.2020.08.002

[9] Z. Teng, X. Yu, and C. Wu, “Iterative attribute augmentation network for face image super resolution,” Electronics Letters, vol. 57, no. 22, pp. 854-856, 2021. https://doi.org/https://DOI:10.1049/ell2.12285

[10] R. J. N. Sharma, “An improved technique for face age progression and enhanced super-resolution with generative adversarial networks,” Wireless Personal Communications: An International Journal, vol. 114, no. 3, pp. 2215-2233, 2020. https://doi.org/https://doi.org/10.1007/s11277-020-07473-1

[11] F. Zhang, J. Zhao, L. Wang, and F. Duan, “3D face model super-resolution based on radial curve estimation. Applied Sciences,” vol. 10, no. 3, pp. 1047-1047, 2020. https://doi.org/10.3390/app10031047

[12] A. B. Deshmukh, and N. U. Rani, “Optimization-driven kernel and deep convolutional neural network for multi-view face video super

resolution," International Journal of Digital Crime and Forensics, vol. 12, no. 3, pp. 77-95, 2020. https://doi.org/10.4018/IJDCF.2020070106

[13] X. Wang, Y. Guo, B. Deng, and J. Zhang, "Lightweight photometric stereo for facial details recovery," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, vol. 13, no. 15, pp. 740-749, 2020. https://doi.org/10.1109/CVPR42600.2020.00082

[14] H. M. R. Afzal, S. Luo, M. K. Afzal, G. Chaudhary, M. Khari, and S. Kumar, "3D face reconstruction from single 2D image using distinctive features," IEEE Access, vol. 8, no. 1, pp. 180681-180689. 2020. https://doi.org/10.1109/ACCESS.2020.3028106

[15] M. Sari, A. Moussaoui, and A. Hadid, "Automated facial expression recognition using deep learning techniques: an overview," International Journal of Informatics and Applied Mathematics, vol. 3, no. 1, pp. 39-53, 2020.

[16] P. Su, "Immersive online biometric authentication algorithm for online guiding based on face recogni3D face reconstruction from single 2D image using distinctive features tion and cloud-based mobile edge computing," Distributed and Parallel Databases, vol. 41, no. 1, pp. 133-154, 2023. https://doi.org/10.1007/s10619-021-07351-0

[17] G. Veselov, A. Tselykh, A. Sharma, and R. Huang, "Applications of artificial intelligence in evolution of smart cities and societies," Informatica, vol. 45, no. 5, 2021. https://doi.org/10.31449/inf.v45i5.360

[18] Y Yang, and X. Song, "Research on face intelligent perception technology integrating deep learning under different illumination intensities. Journal of Computational and Cognitive Engineering," vol. 1, no. 1, pp. 32-36, 2022, https://doi.org/10.14016/j.cnki.1001-9227.2023.04.049