A Research on Two-Stage Facial Occlusion Recognition Algorithm based on CNN

Wang Zhe
Faculty of Engineering and Quantity Surveying, INTI International University, Malaysia | Faculty of Intelligent Engineering, Tianjin Modern Vocational Technology College, China
i21024581@student.newinti.edu.my

Malathy Batumalay
Faculty of Data Science and Information Technology, INTI International University, Malaysia
malathy.batumalay@newinti.edu.my (corresponding author)

Rajermani Thinakaran
Faculty of Data Science and Information Technology, INTI International University, Malaysia
rajermani.thina@newinti.edu.my

Choon Kit Chan
Faculty of Engineering and Quantity Surveying, INTI International University, Malaysia
choonkit.chan@newinti.edu.my

Goh Khang Wen
Faculty of Data Science and Information Technology, INTI International University, Malaysia
khangwen.goh@newinti.edu.my

Zhang Jing Yu
Faculty of Business, Communication and Law, INTI International University, Malaysia
i23024473@student.newinti.edu.my

Li Jian Wei
Faculty of Data Science and Information Technology, INTI International University, Malaysia
i24028610@student.newinti.edu.my

Jeyagopi Raman
Faculty of Engineering and Quantity Surveying, INTI International University, Malaysia
jeyag.raman@newinti.edu.my

Received: 14 August 2024 | Revised: 8 October 2024 | Accepted: 11 October 2024

 DOI: https://doi.org/10.48084/etasr.8736

ABSTRACT

In recent years, pattern recognition has garnered widespread attention, especially in the domain of face recognition. Traditional face recognition methods have certain limitations in unconstrained environments due to factors such as lighting, facial expressions, and poses. Deep learning can be used to address these challenges. This paper proposes a comprehensive approach to face occlusion recognition based on a two-stage Convolutional Neural Network (CNN). Face verification aims at verifying whether two face images belong to the same individual, and it is a more fundamental task compared to face recognition. The process of face recognition essentially involves multiple instances of face verification, sequentially validating different individuals to ultimately determine the corresponding individual for each face. The primary steps in this research include facial detection, image preprocessing, facial landmark localization, facial landmark extraction, feature matching recognition, and 2D image-assisted 3D face reconstruction. A novel two-stage

CNN was designed for facial detection and alignment. The first stage of the network is dedicated to the search for facial windows and regressing vector boundaries. The second stage utilizes 2D images to assist in 3D face reconstruction and perform secondary recognition for cases not identified in the first stage. This method demonstrated excellent performance in handling facial occlusions, achieving high accuracy on datasets such as AFW and FDDB. On the test dataset, face recognition accuracy reached 97.3%, surpassing the original network accuracy of 89.1%. This method outperforms traditional algorithms and general CNN approaches. This study achieved efficient face validation and further handling of unrecognized situations, contributing to the enhancement of face recognition system performance.

Keywords-face detection; two-stage; CNN; occlusion recognition; process innovation

I. INTRODUCTION

With the widespread use of the Internet, face recognition technology has been successfully embedded in major mainstream applications. Compared to human biometrics, such as fingerprints and the iris, face recognition is more convenient to operate, as face information has long-term stability and is easier to collect. Thus, it has been widely used in various application scenarios, including the Internet of Things and medical image recognition [1, 2]. As a branch of computer image processing and pattern recognition research, face recognition technology is currently dominated by deep learning methods [3], which have become the main force promoting the development of artificial intelligence. The latest Convolutional Neural Network (CNN) models have achieved ideal results in face recognition in complex environments. However, there are differences in the effectiveness of face recognition in different scenarios, especially in some special cases. Currently, research has to focus on the impact of the following three scenarios.

The first is the influence of the external environment, including lighting and shooting angles. Light intensity and capturing device characteristics can lead to differences in the RGB values of the images, which can affect the results of subsequent image processing. Image capture equipment generally uses a fixed position, and face information is collected in different scenarios, resulting in different angles of the collected faces and posing a challenge to the richness of sample data. The second is the influence of facial movements since the human face structure is a 3D structure. So, when communicating with the outside world, the facial contour and facial feature images will be different, which can impact the feature extraction of the face recognition algorithm. Finally, there is the impact of face occlusion, which will directly reduce the acquisition of facial information. Masks and other occlusions reduce facial information by 50%, greatly increasing the difficulty of facial recognition. In daily life, jewelry can also cause a certain amount of occlusion, which is a direction that requires optimization.

The proposed facial occlusion recognition system has six primary functional components: facial detection, image preprocessing, facial landmark localization, facial landmark extraction, and feature matching recognition. Figure 1 illustrates the architecture of the proposed system.

A. Face Detection

Face detection is a specific application in the field of object detection [4]. By automatically delineating the area to be detected in the image, the presence of faces in each area is recognized, and if so, the face information is marked. Deep learning algorithms, especially CNNs [5], perform well in complex situations such as different lighting, occlusion, and age changes, laying the foundation for subsequent face recognition.

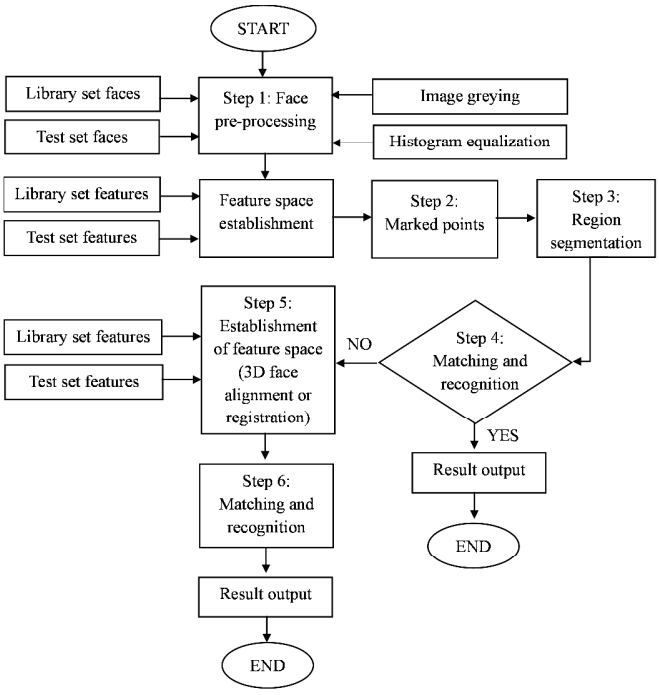

Fig. 1. Flow chart of the proposed system.

B. Image Preprocessing

To improve recognition speed, image preprocessing not only retains the main information of the original image but also eliminates the information that has no impact on the recognition subject. Commonly used image processing methods include grayscale processing, histogram equalization, and median filtering, all based on digital image processing technology.

C. Positioning of Face Landmarks

The goal of face landmark location is to further determine the position of local feature points (such as glasses, eyebrows, nose, mouth, and outer contour of the face) based on determining the general outline of the face. The algorithm compares the relative position relationship of the actual feature points of the face with the standard face template model.

D. Face Feature Point Extraction

Face feature point extraction converts the input face image into a vector representation. By converting the image into a matrix and then scanning and processing it, a result representing the eigenvalues of the region is obtained. Specifically, if an image has 256 colors, then each pixel of the image is a value between 0 and 255, which converts an image into a matrix. This process usually involves counting the processed values of the RGB data for each scan.

E. Feature Comparison and Recognition

In the comparison phase, an exhaustive search is performed on all possible detection windows containing faces, key information is obtained, and then stored or compared. In the comparison process, the similarity comparison of individual parts can be used, merging the results, or directly comparing the global features containing each part and directly outputting the comparison results. This process is relatively time-consuming but is critical to the accuracy of facial recognition.

F. 2D Image-Assisted 3D Face Reconstruction

To further improve the accuracy of the face recognition process, a 3D face reconstruction algorithm based on weakly supervised learning is proposed, and the 3D face reconstruction and optimization of the reconstruction model were enhanced by using the 2D face image to assist in the 3D face reconstruction.

II. METHODOLOGY

In everyday image captures, the distance between the face and the device, different angles, and occlusions all affect the accuracy of face recognition. Specifically, when the face is obscured by masks and other obstructions, the accuracy of the face recognition algorithm decreases significantly. To improve the accuracy of face recognition, this study improved a CNN, focusing on the effective incorporation of an attention mechanism. This improvement focuses on the area above the mouth and nose and reduces the dependence on other regions, thus effectively improving the accuracy of face recognition.

A. Feature Point Calibration Algorithm Based on Cascaded Convolutional Neural Network (CNN)

Face landmark location is important and is widely applied in the fields of automatic face recognition, expression judgment, and face animation synthesis. Due to the influence of various factors, such as posture, expression, occlusion, and lighting, accurate feature point positioning is difficult. Feature points include the eyes, nose, mouth, eyebrows, and borders, which are of great significance for analyzing facial features and identifying identities.

A two-level CNN is proposed to reduce the calibration error, which realizes the localization process from coarse to fine. The sample enrichment strategy was adopted to augment the samples by translation, rotation, random noise, and other methods. In the first layer of the cascade structure, gradient features are introduced, a network model of parallel connection of pixel-domain and gradient-domain is constructed, and the positioning results of the two are weighted and fused. Experiments showed that the calibration error of the improved algorithm was significantly reduced, offering an advantage in detection performance compared to traditional detection algorithms.

The error rate is used to measure the difference between the feature point calibration result and the real coordinates, while the average error rate is used to evaluate the accuracy of the feature point calibration algorithm. At the same time, the average false detection rate is defined as the proportion of all feature points with an error rate of more than 5% to all feature points, which is used as the standard deviation of the detection results. These two indicators are selected to evaluate the feature point calibration algorithm. The error rate for measuring the error between the feature point calibration result and the real coordinates is given as:

$$ErrorRate_J = \frac{\frac{1}{M}\sum^{M}|p_{ij}-q_{ij}|}{I_j} \quad (1)$$

where $ErrorRate_j$ is the error rate of the j^{th} image and M denotes the number of feature points in the face. The measurement function $|p_{ij} - q_{ij}|$ is calculated using:

$$\left|p_{ij} - q_{i,j}\right| = \sqrt{(P_x - q_x)^2 + (p_y - q_y)^2} \quad (2)$$

After obtaining the error rate of a sample, the average error rate of all samples is calculated as:

$$ErrorRate = \frac{1}{N}\sum_{f=1}^{N} ErrorRate_f \quad (3)$$

where N indicates the total number of samples in the dataset. The average error rate can be used to measure the accuracy of the feature point calibration algorithm. At the same time, if the error rate of a feature point is greater than 5%, it is assumed that the point is falsely detected, and the average $FailureRate$ is given by:

$$FailureRate = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left\{\frac{|p_{ij}-q_{ij}|}{I_i} > 0.05\right\}}{M \cdot N} \quad (4)$$

The average false detection rate can be used to measure the proportion of all feature points with an error rate of more than 5% in the test sample set, which can be regarded as the standard deviation of the detection results of the feature point calibration algorithm. This study selected the average error rate and the average false detection rate to evaluate the feature point calibration algorithm.

B. Face Verification Based on Deep Learning

The face detection algorithm of the bi-hierarchical CNN adopted in this study realizes the effective detection and localization of the face in the original image by constructing an image pyramid, a sliding window, a cascaded CNN, and a correction network. The detailed steps of the process are as follows:

1. Build image pyramid: Use a sliding window 40×40 in size and 4 steps to build an image pyramid.
2. Slide the window to traverse the original image: Slide through the original image, zoom the window image to 12×12, and send it to 12-net for face judgment, where the judgment threshold is $ThrFace1 = 0.95$.

3. 12-calibration-net correction network: Correct the area identified as the face in step 2 and adjust the size and position of the face frame.
4. Non-maximum suppression: Eliminate redundant faces and retain high-confidence face frames.
5. Zoom to 24×29 for face judgment: Zoom the filtered face frame to 24×29 and send it to 24-net for face judgment to further eliminate the false detection area. The judgment threshold is *ThrFace*2 = 99.
6. 24-calibration-net correction network: The remaining face frames in step 5 are corrected, the redundant faces are eliminated by using a non-maximum suppression algorithm, and the screened faces are represented by a center point.
7. Image scaling: Scale the length and width of the image by 1.2 times and return to step 2 if the image size is greater than 40×40.
8. Deduplication strategy: Utilize a deduplication strategy to eliminate orphaned candidates that appear in all layers of the pyramid.
9. The cascaded CNN face detection algorithm ends.

C. 2D Image-Assisted 3D Face Reconstruction

The 3D face reconstruction model based on weakly supervised learning consists of the following three main modules.

1) CNN-Based Regressor

The CNN-based regressor is used to estimate the 3DMM (3D Morphable Model) parameters of the input 2D image. This regressor model processes the 2D images with 3D annotated data and the face images of natural scenes. The predicted 3DMM coefficients α_j and α_i are output. In this study, ResNet-50 is mainly used to extract the features of the input 2D face image, and the last two layers of the network are replaced by two Fully Connected (FC) layers. Finally, a 62-dimensional 3DMM parameter vector is output. Figure 2 shows the network structure.

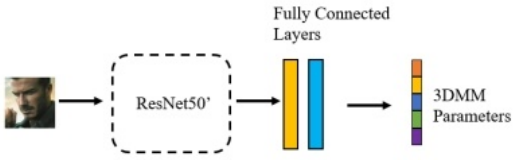

Fig. 2. Regressor module.

2) Encoder

The encoder is used to extract high-dimensional features from an input image. This module performs feature extraction on the input image and represents it as a high-dimensional feature vector. In this module, six convolutional layers are used, each followed by a ReLU activation function and a maximum pooling layer. In addition, the 3DMM coefficient conditional adversarial network module is mainly composed of four FC layers, and then a softmax layer is added to output the results of the consistency judgment of the network to the input pair, as shown in Figure 3.

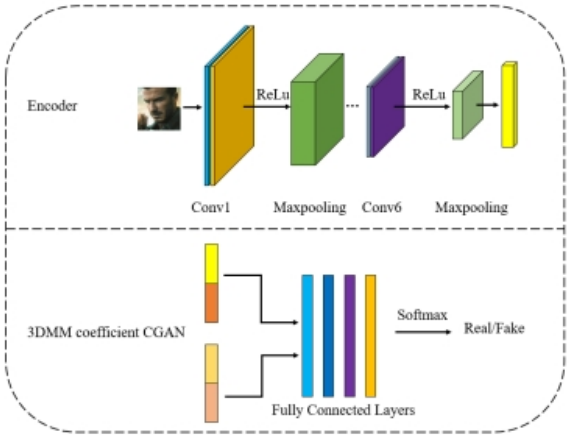

Fig. 3. Encoder and conditional GANs module.

3) 3DMM Coefficient Conditional Adversarial Network

This network is used to evaluate whether a pair of inputs (high-dimensional features, 3DMM coefficients) match. The network is trained on two paths: the top is used to reconstruct 3D faces, and the bottom uses conditional GANs based on 3DMM coefficients for weakly supervised learning. This helps to optimize the reconstruction accuracy of the 3D face model and the robustness of images in natural scenes.

According to the design of the above three modules, the model learns by minimizing the traditional 3D face loss function and a 3DMM coefficient conditional adversarial network ϱ_{3d} loss function. The first loss function adopts the asymmetric Euclidean loss function in the Robust 3DMM algorithm, which is used to measure the accuracy of the model in predicting the 3DMM coefficient. The second loss function ϱ_{cgan} is called the 3DMM coefficient condition against the network loss, under the condition of knowing the potential representation of the face used to evaluate whether the predicted 3DMM coefficient conforms to the distribution of the real data to improve the accuracy of the model to reconstruct the 3D face. To differentially learn for different tasks, the weight coefficient λ ($\lambda = 0.005$) is added to the loss function. Therefore, the loss function of the whole model is:

$$\varrho = \varrho_{3d} + \lambda_{cgan}^{\iota} \quad (5)$$

Overall, by combining these three modules, the model can estimate 3DMM parameters from the input 2D images and optimize them through a conditional adversarial network to improve the robustness of natural scene images and the accuracy of 3D face reconstruction.

D. Hardware Environment

When using deep learning algorithms to process image data, due to the complex network structure, the huge number of parameters, and the large increase in the number of training samples, the limitations CPUs in processing computationally

intensive data are becoming more and more obvious. Thus, more and more researchers choose GPUs as the main computing core to address the shortcomings of slow training speed. As a result, researchers can use a larger training set and a deeper and more complex network to extract the deep features of the sample. Table I shows the main hardware resources and systems used in the experiments.

TABLE I. PC CONFIGURATION

Operating system	**Ubuntu16.04**
CPU	Intel Core i7-4790 @ 3.60 GHz
GPU	NVIDIA GeForce GTX1080, 8GB VRAM
Memory	16 GB

E. Datasets

This study used two datasets for model validation.

1) LFW

In 2007, the Computer Vision Laboratory at the University of Massachusetts Amherst in the United States completed the collation of the Labeled Faces in the Wild (LFW) dataset. This is an unconstrained natural scene face recognition dataset used to study the face recognition problem in unrestricted situations, which is composed of 13,323 face pictures of celebrities around the world in different orientations, expressions, and lighting environments of natural scenes on the Internet, on a total of more than 5,749 people. Among them, 4,069 people had only one face picture, while the remaining 1,680 people had more face pictures. Each face picture is distinguished by its unique name ID and serial number. This ensemble is widely used to evaluate the performance of face verification algorithms and has become the most commonly used dataset in academia to evaluate the performance of face verification.

The original images of the LFW dataset were obtained by grabbing the news image and the corresponding headline information from the Yahoo News channel, followed by face detection, image deduplication, face labeling, cropping, and size normalization, and forming a training test set.

For the division in training and test sets, assuming that the accuracy is p_i, the final evaluation index is the average accuracy $\hat{\mu}$ and the mean standard deviation $\hat{\sigma}$, given by:

$$\hat{\mu} = \frac{\sum_{i=1}^{10} p_i}{10} \quad (6)$$

$$S_E = \frac{\hat{\sigma}}{\sqrt{10}} \quad (7)$$

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{10} (p_i - \hat{\mu})^2}{9}} \quad (8)$$

2) CASIA-Webface

In 2014, the Institute of Automation of the Chinese Academy of Sciences used a semi-automated method to collect and establish the CASIA-Webface dataset from the Internet [6]. The dataset contains 10,575 people with a total of 494,414 face images from the IMDB website. IMDB is a movie database website with a large number of pictures of movie stars, and the dataset contains images from each star's personal album. After the images were collected, a series of automatic cleaning and manual checks were performed to ensure the correctness of the dataset labeling. As the LFW dataset is the most commonly used dataset for face verification, Casia-Webface excludes the star faces that appear there.

III. RESULTS AND DISCUSSION

After the feature extraction of the sliding window CNN, 25 320-dimensional feature vectors were obtained. After feature fusion, a 25×320=8,000-dimensional feature vector was obtained to represent the face. This feature vector is then mapped to a lower-dimensional space by dimensionality reduction, the feature information of the face is extracted, and finally, the face verification classifier is used for face verification. The experiments showed that the proposed network model had a higher accuracy on the LFW dataset after PCA dimensionality reduction processing.

The PCA dimensionality reduction method was used to reduce the dimensionality of the fused features. The best accuracy of the model on LFW was achieved when the fused features were not processed and dimensionality was reduced. The results in Table II show that the proposed model can achieve higher accuracy when PCA dimensionality reduction changes to a moderate dimension (150-300). Reducing too much or too little dimensions leads to a decrease in accuracy.

TABLE II. PCA DIMENSIONALITY REDUCTION

Dimensionality reduction	**LFW accuracy**
PCA 50	0.9788
PCA 100	0.9804
PCA 150	0.9848
PCA 200	0.9873
PCA 250	0.9844
PCA300	0.9832
PCA 350	0.9822
PCA 400	0.9810

A. Judgment of Feature Similarity

As mentioned above, two face images will be respectively given the feature vectors of the corresponding two face images after the same CNN structure and the same network parameters. Comparing whether the two faces represent the same person can be judged using Euclidean distance, cosine distance, or joint Bayesian methods. These three distance measurement methods were compared. For each method, model accuracy was measured three times, and then the average value was obtained. This average value was used to represent the accuracy of the model, as shown in Table III. The joint Bayesian method was more effective than the direct use of Euclidean distance in the face verification task. As the joint Bayesian method uses the class information in the training data, the results were better compared to the other two methods.

TABLE III. COMPARISON OF SIMILARITY EFFECTS

Categorization	**Accuracy 1**	**Accuracy 2**	**Accuracy 3**	**Average accuracy**
Continental distances	0.9809	0.9823	0.9819	0.9817
Cosine distance	0.9801	0.9815	0.9807	0.9808
Joint Bayes	0.9825	0.9836	0.9832	0.9831

B. Activation Functions

Many nonlinear activation functions have been proposed for deep neural networks, and an activation function can produce different effects in different network models, which are related not only to the network structure but also to the task being processed and the input data. This study selected four functions for comparison, namely ReLU, PReLU, RreLU, and ELU. Except for the activation functions, the other parameters of the experiments were the same to exclude interference from other factors. Table IV shows the experimental results.

TABLE IV. COMPARISON OF ACTIVATION FUNCTION EFFECTS

Numbering	Activation function	LFW accuracy
1	ReLU	0.9773
2	PReLU	0.9812
3	RReLU	0.9841
4	ELU(α=0.25	0.9832
5	ELU (a =0.5)	0.9829
6	ELU (α =1)	0.9813

The results show that RReLu performed better than the other activation functions. Parameter a of RReLU was randomly selected from the Gaussian distribution and changed accordingly with the network training. This can suppress the overfitting phenomenon that may occur due to too many network parameters.

C. Training Data

The rapid development of deep learning is inseparable from the support of data, and training data has a crucial impact on model results. The amount of data generally affects the performance of the model, including accuracy, whether it is overfitted, training time, etc. Data quality is the most important factor affecting the model, and high-quality data can make the model achieve very good results. In the preprocessing phase, image data were typically enriched in a variety of ways, such as mirror flipping, rotating, cropping, distorting, and adding salt-and-pepper noise. The results in Table V are shown in terms of the quantity, angle change, size change, and color of the training data.

TABLE V. COMPARISON OF TRAINING DATA EFFECTS

Description of experiment	Description of training data	LFW accuracy
Different numbers of face image samples used for training	5000 people	0.9778
	7000 people	0.9783
	9000 people	0.9803
	10575 people	0.9815
	10,575 people, randomly flipped	0.9834
	10,575 people, random flip, grayscale map	0.9802
10,575 people randomly cropped and mirrored	No random cropping, no image expansion	0.9793
	It is not randomly clipped and the image is expanded	0.9819
	Random cropping, no image expansion	0.9837
	Random cropping, image expansion	0.9845

D. Optimal Combination of Parameters

After comparing the above conditions, the optimal combination for each condition was selected as follows: The feature vector obtained by the sliding window CNN was reduced to 200 dimensions by PCA, the joint Bayesian method was used to judge the recognition of faces, RReLU was selected as the optimal activation function, and, finally, the face images of 10,575 people were selected, which were randomly flipped, randomly cropped, and mirrored.

1) Experiments Based on DeepID

In this experiment, face recognition based on the DeepID network was performed on the LFW face database. The DeepID network structure was used on the Tensorflow deep learning platform, and the model training adopted adaptive gradient descent with a learning rate of 0.0001, a momentum of 0.9, a maximum of 10,000 iterations, and a batch (mini_batodh) of 40. The training results in Figure 4 were obtained, with an accuracy of 95.94% and a minimum loss of 0.2472.

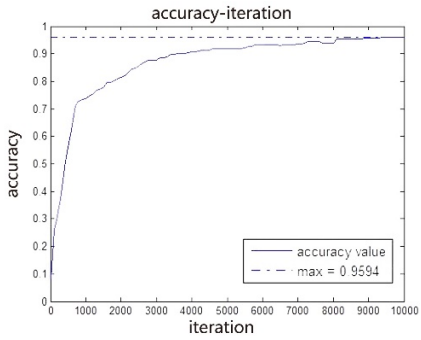

Fig. 4. DeepID model training accuracy.

2) Experiments Based on the Basic Network Structure of the Proposed Method

This experiment was based on the basic network structure, and the input data were RGB images with a size of 224×224. The stochastic gradient descent training algorithm was adopted, the basic learning rate was 0.0001, which was reduced by 0.9 times every 300 iterations, the maximum number of iterations was 20,000 times, the size of a batch (mini_batch) was 40 images, and the momentum was 0.9. The training results, shown in Figure 5, indicate that the validation accuracy of the model reached 97%, and the minimum loss reached 0.0008298.

3) Experiments Based on the Proposed Optimized Model

This experiment used the proposed optimized face recognition model, based on a CNN with residual module and a sliding window. The training algorithm used stochastic gradient descent, the basic learning rate was 0.0001, which was reduced by 0.9 times every 300 iterations, the maximum number of iterations was 10000 times, the size of a batch (mini_batch) was 40 images, and the momentum was 0.9. The training results in Figure 6 show that the validation accuracy reached 98.75%, and the minimum loss reached 0.01695. For comparison, the accuracy was 2.81% higher than that of DeepID and 1.75% higher than that of the basic network.

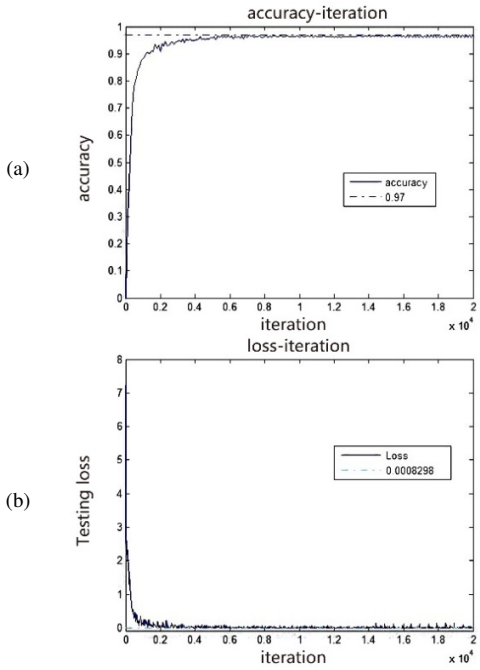

Fig. 5. Training accuracy (a) and loss (b) of the basic network model.

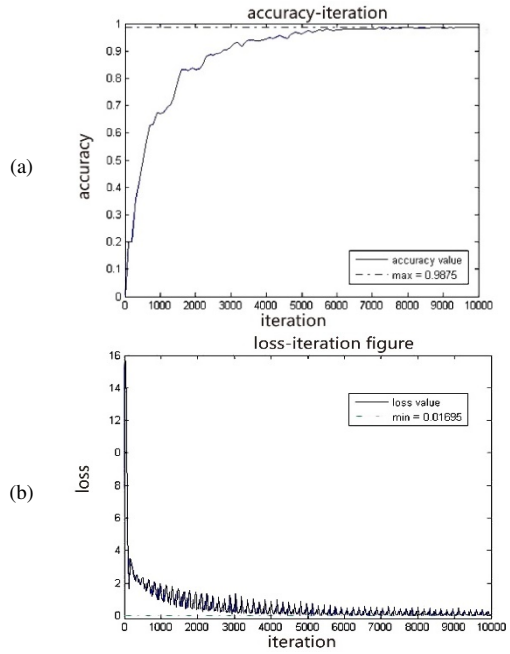

Fig. 6. Optimized model's training accuracy (a) and loss (b).

4) Comparison and Analysis

The accuracy of the basic model was 97% on the LFW face database, while the accuracy of the proposed neural network based on residual network and sliding window was 98.75%, indicating an improvement of 1.75%, as shown in Table VI. Several scientific research institutions and technology companies have successively proposed various face recognition models. This study compared some deep learning-based face recognition models, including DeepFace [7], DeepID, DeepID2, DeepID2+ [8], DeepID3 [9], VGG [10], and FaceNet [11]. The accuracy of the optimized algorithm was higher than that of DeepFace's multimodel, DeepID's multimodel, DeepID2's single model, and DeepID2+'s single model, and only lower than that of the FaceNet model, DeepID2+ multimodel, and DeepID2 multimodel developed by Google. Considering single models, the proposed optimized model had lower accuracy only compared to the FaceNet model, which was trained on a dataset of 200 million images. However, it is difficult for general scientific research institutions to have such huge data and computing resources to conduct experiments.

TABLE VI. COMPARISON OF SIMILARITY EFFECTS

No	Model	Number of models	Number of training images	Average recognition accuracy
1	DeepFace	1	About 4 million	95.92%
2	DeepFace	7	About 4 million	97.35%
3	DeepID	1	About 200,000	91.54%
4	DeepID	60	About 200,000	97.45%
5	DeepID2	1	About 200,000	95.43%
6	DeepID2	25	About 200,000	98.97%
7	DeepID2+	25	About 290,000	98.97%
8	DeepID3	25	About 300,000	99.53%
9	VGG	1	About 2.5 million	97.27%
10	FaceNet	1	About 200 million	99.63%
11	Proposed	1	About 520,000	98.75%

In this study, 68 key face points and dense 2D or 3D face coordinate information were used. From another point of view, 68 face key points can be regarded as the sampling of dense face point cloud data. Since PRNet [12] and VRN-Guided [13] were not based on the 3DMM method, and the point cloud data generated by these two methods did not correspond to the 3DMM, they were only compared to the sparse 68 face key points. The results in Figure 8 show the Error Distribution Curves (EDC) of the proposed algorithm in the AFLW2000-3D dataset, where the horizontal direction in the coordinate system represents the increase in NME%, the vertical direction represents the number of images tested (the maximum value is 1980), and the average NME% of each method is shown in the lower right corner. As can be observed, the proposed model (2DASL) achieved the lowest NME% among all others in the evaluation of the error between the 2D and 3D coordinates of the face and the real annotated information. On the other hand, for 3DMM-based methods, such as 3DDFA [14] and DeFA [15], the proposed method achieved better performance than the above methods in 68 key face point alignment and dense face alignment tasks.

Studying the performance of the proposed method in different face poses and datasets, it was observed that it corresponded well to AFLW2000 small face angles. The NME values measured at medium and large angles were also calculated as the average NME in the AFLW2000-3D and AFLW-LPFA datasets, as shown in Table VII. The proposed optimized method achieved the lowest NME values on both test datasets and the lowest NME values in each face pose test in the AFLW2000-3D dataset, even surpassing the current best PRNet algorithm, reducing the NME values by 0.009 and 0.008 in the AFLW2000-3D and AFLW-LFPA datasets, respectively. Especially in the large-posture face image set (from 600 to

900), the NME of the optimized method was 0.2 lower than that of PRNet. Therefore, it can be assumed that the proposed 3D face model can achieve better performance using the 2D face image in the natural scene to participate in the training process.

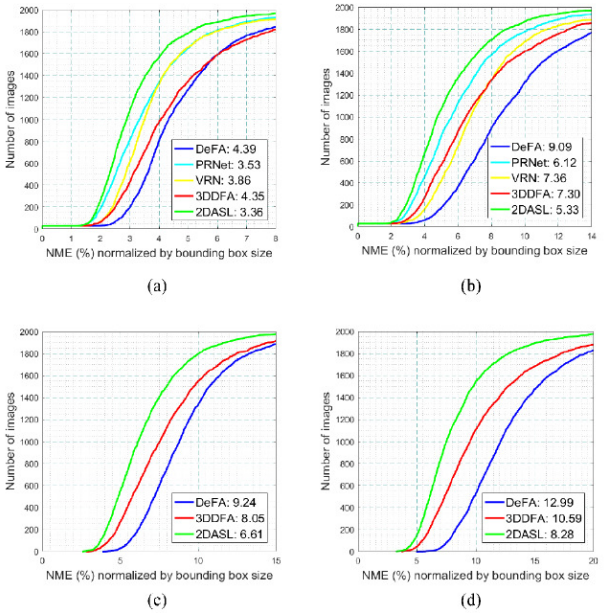

Fig. 7. Face alignment error distribution curve. (a) 68 2D face key point alignment tests, (b) 68 3D face key point alignment tests, (c) dense 2D face alignment, (d) Dense three-dimensional face alignment.

TABLE VII. DIFFERENT METHODS IN AFLW2000-3D (68PTS) AND AFLW-LFPA (34PTS)

Method	**AFLW2000-3D Dataset (68pts)**				**AFLW-LFPA (34pts)**
	[0°,30°]	[30°,60°]	[60°,90°]	Mean	Mean
3DDFA	3.78	4.54	7.93	5.42	
DeFA				4.50	3.86
PRNet	2.75	3.51	4.61	3.62	2.93
Proposed	2.75	3.44	4.41	3.53	2.85

IV. CONCLUSIONS

This study presented a comprehensive approach to facial occlusion recognition based on a two-stage CNN. The primary steps in the proposed method included facial detection, image preprocessing, facial landmark localization, facial landmark extraction, feature matching recognition, and 2D image-assisted 3D facial reconstruction. A novel two-stage CNN was designed for facial detection and alignment. The first stage of the network was dedicated to the search for facial windows and regressing vector boundaries. The second stage utilized 2D images to assist in 3D face reconstruction and perform secondary recognition for cases not identified in the first stage. This method demonstrated excellent performance in the handling of facial occlusions, achieving a testing accuracy of 97.3%.

Future studies could investigate enhancing the algorithm's computational efficiency for real-time applications and augmenting the dataset with more annotated 2D-3D facial images to further increase model resilience. Exploring alternative deep learning models, particularly transformer-based architectures, for feature extraction and 3D reconstruction may produce better results. Furthermore, extending the method's usage across diverse face recognition contexts, including security and biometric systems, will enhance its validation and enhance its practical significance.

REFERENCES

[1] A. M. Al-Ghaili *et al.*, "A Review: Image Processing Techniques' Roles towards Energy-Efficient and Secure IoT," *Applied Sciences*, vol. 13, no. 4, Jan. 2023, Art. no. 2098, https://doi.org/10.3390/app13042098.

[2] A. Halder, S. Gharami, P. Sadhu, P. K. Singh, M. Woźniak, and M. F. Ijaz, "Implementing vision transformer for classifying 2D biomedical images," *Scientific Reports*, vol. 14, no. 1, May 2024, Art. no. 12567, https://doi.org/10.1038/s41598-024-63094-9.

[3] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017, https://doi.org/10.1016/j.media.2017.07.005.

[4] B. Ye *et. al.*, "Small Target Detection Method Based on Morphology Top-Hat Operator," *Journal of Image and Graphics*, 2002.

[5] L. Chang *et al.*, "Convolutional neural networks in image understanding," *Acta Automatica Sinica*, vol. 42, no. 9, pp. 1300–1312, Sep. 2016, https://doi.org/10.16383/j.aas.2016.c150800.

[6] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, Jun. 2014, pp. 1701–1708, https://doi.org/10.1109/CVPR.2014.220.

[7] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 2892–2900, https://doi.org/10.1109/CVPR.2015.7298907.

[8] Y. Sun, D. Liang, X. Wang, and X. Tang, "DeepID3: Face Recognition with Very Deep Neural Networks." arXiv, Feb. 03, 2015, https://doi.org/10.48550/arXiv.1502.00873.

[9] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC 2015 - Proceedings of the British Machine Vision Conference 2015*, 2015.

[10] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 815–823, https://doi.org/10.1109/CVPR.2015.7298682.

[11] V. Blanz and T. Vetter, "A Morphable Model For The Synthesis Of 3D Faces," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 1st ed., M. C. Whitton, Ed. New York, NY, USA: ACM, 2023, pp. 157–164.

[12] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, and R. Chellappa, "UMDFaces: An annotated face dataset for training deep networks," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, Denver, CO, USA, Oct. 2017, pp. 464–473, https://doi.org/10.1109/BTAS.2017.8272731.

[13] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos, "Large Pose 3D Face Reconstruction from a Single Image via Direct Volumetric CNN Regression," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, Oct. 2017, pp. 1031–1039, https://doi.org/10.1109/ICCV.2017.117.

[14] Xiangyu Zhu, Junjie Yan, Dong Yi, Zhen Lei, and S. Z. Li, "Discriminative 3D morphable model fitting," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Ljubljana, Slovenia May 2015, pp. 1–8, https://doi.org/10.1109/FG.2015.7163096.

[15] Y. Liu, A. Jourabloo, W. Ren, and X. Liu, "Dense Face Alignment," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Venice, Italy, Oct. 2017, pp. 1619–1628, https://doi.org/10.1109/ICCVW.2017.190.