ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20240824001940
Original scientific paper

A Deep Learning-Based Framework for Robust Facial Keypoint Localization in Unconstrained Conditions

Li Juan YANG*, Ying LI

Abstract: Facial keypoint localization plays a critical role in facial recognition, security monitoring, and human-computer interaction. Traditional methods rely heavily on handcrafted features, making them sensitive to occlusions, lighting variations, and pose changes. This study proposes a deep learning-based framework integrating lightweight convolutional neural networks (CNNs) and Conditional Random Fields (CRFs) to improve keypoint detection and localization accuracy under unconstrained conditions. A fast connected convolutional layer is introduced in a cascaded network structure, significantly reducing feature space information loss and enhancing geometric relationship modeling. The results showed that the proposed face detection model had a small cumulative error value, with a feature recognition accuracy of over 0.9, and an average accuracy of over 90 for all classes under three different image conditions. The proposed localization model had smaller error values and a much lower error rate than other algorithms under various segmented image differences, effectively considering data feature differences and achieving higher localization accuracy. The proposed deep learning model can effectively achieve the fusion of output features and improve the effectiveness of facial keypoint localization.

Keywords: face; conditional random fields; convolution; convolutional neural network; deep learning; keypoint

1 INTRODUCTION

The advancement of science and technology, as well as the improvement of computing power, have led to the widespread application of facial detection and recognition technology in fields such as security monitoring, identity verification, financial prevention and control, and human-computer interaction. Moreover, the demand for this detection technology has also expanded to areas such as digital video processing and content retrieval [1]. As the core content of biometric recognition, facial recognition technology needs to identify key information of obvious and easily distinguishable features in the face. Facial detection covers a wide range of content, including posture, position coordinates, and shape and size. The challenge of this technology lies in the recognition of subtle facial differences, such as environmental lighting, facial expressions, line of sight occlusion, and other factors that can affect the distance between facial features, thereby affecting the recognition effect and accuracy [2]. Facial recognition is a common and complex visual model, and the visual signals it conveys play an important role in interpersonal communication and interaction. Facial processing has a wide range of application fields and development prospects, including facial recognition, facial tracking, and other aspects. Facial detection is a prerequisite for all processing steps, and it is commonly used in image databases to distinguish between faces and backgrounds in frame images for image processing [3]. Facial detection requires determining the pose, position coordinates, shape and size of the detected face. Although the facial structure composed of organs such as eyebrows, eyes, nose, and mouth is relatively determined, the difficulty of facial detection is still high due to changes in expression, pose, and occlusion of light and shadows. Facial recognition schemes typically locate the eye area and extract overall features from windows centered around each region of interest [4]. The coordinates of the key points located are related to geometric attributes such as sea distance and angle. In fact, facial models for human body measurement often combine information sources, configuration sources, and appearance sources, and graph based recognition methods have been proven to have good application effectiveness. The purpose of detecting facial keypoint information is to ensure that all identified faces correspond to their positions, and the output result is information such as the bounding rectangle of the face and the coordinates of each keypoint in the image [5]. Deep learning methods have brought significant improvements in fields such as image recognition and analysis, speech recognition, and natural language processing in the past. Facial keypoint localization, which identifies and annotates the accurate positions of key facial features such as eyes, nose, and mouth through technological means, is the foundation for achieving efficient facial recognition. This technology is not only crucial for improving the accuracy of facial recognition, but also has significant implications for the application of facial expression recognition, dynamic facial tracking, and even virtual reality technologies. Deep learning techniques can achieve algorithm adaptability based on network structure design when processing image feature information. Probabilistic graph models have been applied in current computer vision tasks and can effectively analyze the structural relationships of random variables. However, traditional deep learning methods are difficult to capture the structural relationships between facial features, and their application effectiveness is limited under the influence of factors such as posture changes, facial occlusion, and expression changes [6]. Therefore, in order to significantly improve the accuracy of facial keypoint localization and model generalization ability in non standardized environments, a facial keypoint detection and localization algorithm based on deep learning theory is proposed. This algorithm introduces the Conditional Random Field (CRF) and combines it with a Deep Learning (DL) network to predict the structural changes of facial landmark positions while maintaining the geometric relationships of keypoints, completing the detection of facial keypoints. CRF, as a discriminative probability model, effectively absorbs the advantages of maximum entropy model and hidden Markov model. Its layout structure has good applicability

in algorithm design and is commonly used in fields such as natural language sequence analysis and biological text annotation [7]. At the same time, considering the problem of keypoint localization under unconstrained conditions, the research proposes the idea of using convolutional neural networks (CNN) to achieve the fusion of facial feature information, and improves on data samples, depth features, and loss functions to enhance feature mining capabilities and achieve information convolution operations. The research on facial keypoint localization technology aims to provide reference for the development and universal application of the entire facial recognition technology.

2 LITERATURE REVIEW

FKPD technology is a localization technique for facial information and contour peripheral landmarks. With the development of computer vision technology and artificial intelligence, algorithms related to face detection have gradually evolved from early template matching techniques and ensemble learning frameworks to the DL stage. Yu et al. conducted a literature review on DL-based facial anti-spoofing technology and concluded that the application of multi-mode or dedicated sensors has great development prospects for anti spoofing problems in facial recognition [8]. Bisogni et al. proposed the idea of combining pose estimation and feature matching, using fractal coding to extract facial feature vectors. This method effectively improved the efficiency of facial recognition [9]. Wen et al. designed a lightweight cascaded facial landmark detector with adaptive computation and improved the network architecture allocation efficiency through group search algorithm. This method effectively reduced the position loss information of predicted keypoints [10]. Gupta et al. proposed a 2D facial image recognition method based on a combination of accelerated robust features and scale invariant feature transformations. The maximum recognition accuracy demonstrated by this method exceeded 98% [11]. Ahila Priyadharshini et al. proposed a six layer Deep CNN (DCNN) architecture for facial ear information recognition. The recognition rate of this deep model on the dataset exceeded 95%, and it still showed good robustness in different environmental tests [12]. Vu et al. proposed a combination of DL and local binary patterns to address the challenge of masked face recognition, and designed an encoder based on the Retinex algorithm. This method showed good recognition performance on the dataset, with an F1 score better than other methods [13]. Hariri et al. proposed a masked face recognition method based on occlusion removal and DL features, which utilizes DCNN and multi-layer perceptron to achieve feature extraction and information classification. This method exhibited good recognition performance on real datasets [14]. Misael Burruel Zazueta et al. proposed using FaceNet pre trained network to achieve facial recognition and designed an anti spoofing network architecture. The results indicate that this method has high accuracy in recognition under different lighting conditions and hardware environments [15]. Niu et al. studied the facial key point sequences with the help of point-transform network and mixer attention mechanism to better predict the severity of depression. The results showed that the method can effectively analyse the temporal characteristics of the sequences and has better application [16].

The dataset may have biases in facial expression recognition algorithms due to differences in collection conditions and cultural motivations. Therefore, Li et al. proposed using deep emotion conditional adaptation networks to achieve deep exploration of data features, that is, using the underlying label information of the dataset to complete conditional matching. This method could effectively solve the expression class distribution bias of data and had good potential for application in facial expression recognition tasks in databases [17]. Cowen et al. used DCNN to study human facial expressions and found that there are fine-grained differences in human expressions in different regions [18]. Harakannanavar et al. utilized improved directed gradient histograms, local binary patterns, and fast keypoint detectors to achieve multi feature vector extraction and development for facial expression recognition systems. The results showed that the proposed model exhibited good classification performance of over 95%, and the feature fusion effect was significant [19]. Zhang et al. utilized end-to-end deep learning models to achieve facial image synthesis and pose expression recognition, and utilized generative adversarial networks to locate facial pose expressions. The results indicate that this method can continuously and automatically generate facial images with different expressions and poses, and the application effect is significant [20]. Lv et al. analyzed the 3D facial similarity measurement framework based on Kendall's shape space theory and expressed facial shape features using facial feature landmark models. The recognition accuracy of this method in public facial databases was greater than 97% [21]. Zhang et al. used multi-column CNN to estimate gaze point images, refined the depth of facial image keypoints, and globally optimized them with the captured original depth. This method performed well on datasets [22]. Liu et al. proposed using Point Adversarial Self Mining (PASM) to improve the accuracy of facial expression recognition, by simulating the human learning process to correlate the information correlation between localization and target tasks. The results show that the method exhibits good application performance [23]. Dar et al. processed facial reconstruction images using deep learning methods, utilizing a novel supervised deep learning technique - stacked deep autoencoder - to complete image processing, modeling, and reconstruction. The results show that this method can effectively reconstruct the original facial image, with a classification error value of only 0.1, which is much smaller than other algorithms [24]. Considering the impact of abnormal noise points on portrait segmentation errors, Wu et al. proposed a sparrow search clustering algorithm based on wavelet proportional shrinkage and improvement to address the challenges of frontal portrait segmentation and used elite inverse learning methods and adaptive weighting factors to accelerate the convergence performance of K-Means algorithm. The results indicate that this method has higher image segmentation accuracy than traditional methods and can adapt to different lighting conditions [25].

In summary, the advantages of DL in current image processing and computer world processing have attracted most scholars to apply it in facial keypoint localization and recognition. When increasing the depth of CNN, it can lead to the problem of excessive expansion. Previous scholars have often used the form of adding difference networks to increase the network depth. However, this method makes it easier for the information in the feature space under network overlay to be lost, leading to the problem of keypoint pixel regression. Based on this, this study takes into account the difficulty of facial keypoint localization. In addition to considering computational complexity and modeling quality, the paper also improves the Memory Access Cost (MAC) and Computational Energy Efficiency Parameters (CEEP) of the network to achieve improved keypoint detection accuracy. Subsequently, an improved CNN is introduced to achieve the same sum of feature information through regression cascading to locate key points.

3 METHODOLOGY AND EXPERIMENTAL SETUP

This study leverages the advantages of CNN and proposes to combine lightweight networks with mathematical models in Part 1, and to improve cascaded networks in Part 2 to enhance facial recognition accuracy.

3.1 Facial Keypoint Detection Based on DL

CRF determines that nodes are only related to neighboring nodes and is an output sequence conditional probability distribution model formed when a set of inputs is fixed. The CRF fully connected form can connect any pixel point. It can combine and adjust the classification results of DL with the original image elements, effectively optimizing the fuzzy and uncertain areas in the classification results, thereby clarifying the edges and correcting misclassified results [26]. When the variables in CRF have the same structure, the conditional probability of the linear chain CRF can be expressed as Eq. (1).

$$P\left(y|x\right)= \frac{1}{Z\left(x\right)}\exp\left\{\sum_{i,k}\lambda_k t_k\left(y_i-1,y_i,x,i\right)+\sum_{i,l}\mu_l s_l\left(y_i,x,i\right)\right\} \quad (1)$$

In Eq. (1), x is the given variable. y is the output variable. i is the number of variables. t_k is the transition feature, and s_l is the state feature. k and l are the number of corresponding features. $Z(x)$ is the normalization factor. λ_k and μ_l are connection weights. CRF defines feature functions based on the empirical characteristics of the data when processing it. This study replaces binary Energy Function (EF) with ternary EF to output mean and accuracy matrix in fully connected CRF. The conditional probability calculation formula is Eq. (2).

In Eq. (2), θ_1 is the neural network parameter. $Z_\theta(T)$ is the partition function. C_{ij} is a symmetric positive definite matrix. Y_i and Y_j are facial keypoints. $\Phi_{\theta_1}\left(Y_i,T\right)$ is a unary EF. (Y_i, Y_j, ξ) is a ternary EF. T is the facial image. N is the number of key points. ξ is a deformable model parameter.

$$p_\theta\left(Y,\xi|T\right)= \frac{1}{Z_\theta(T)}\exp\left\{-\sum_{i=1}^{N}\Phi_{\theta_1}\left(Y_i,T\right)-\sum_{i=1}^{N}\sum_{j=j+1}^{N}\varphi_{C_{ij}}\left(Y_i,Y_j,\xi\right)\right. \quad (2)$$

This study introduces Cholesky decomposition to solve the speed impact problem of traditional linear solving, which converts the accuracy matrix into a lower triangular matrix and its transpose form, and takes diagonal matrix values for rows and columns. Currently, most scholars focus on the convolutional bottleneck structure and modeling size in the design of efficient networks, while neglecting other factors that affect time consumption [27]. In addition to considering computational complexity and modeling quality, this study also analyzed MAC and CEEP. In DL, when there are significant changes in network characteristics, the energy consumption of memory access may exceed the computational load. The MAC formula for the convolutional layer is Eq. (3).

$$\mathrm{MAC}=hw\left(c_i+c_0\right)+k^2c_ic_0 \quad (3)$$

In Eq. (3), h represents high feature. w is the characteristic width. k is the size of the convolution kernel. c_i and c_0 are the number of input and output channels. When the expression of the convolutional layer remains unchanged, the mean inequality can be used to determine that the efficiency of the algorithm can only be guaranteed when the number of channels in both the input and output layers is equal. The different receptive fields of the VoVNet model are concatenated to aggregate shallow features, which can reduce the number of model parameters while preserving the original information of the data. The aggregation module can complete feature map processing through a One-Shot Aggregation (OSA) strategy, which can effectively ensure the expansion of the receptive field while avoiding the problem of resource and computational waste caused by repeated convolutions [28]. The foundation of the VoVNet model is the Densely Connected Convolutional Networks (DenseNet) model. The DenseNet architecture connects the convolutional network layers in a feedforward manner. Its dense connection form allows any layer to take the output information of the previous layer as its own input content, thereby ensuring the maximization of information flow between network layers. Eq. (4) is the input expression of DenseNet.

$$x_i=H\left(\left[x_0,x_1,x_2,\ldots,x_{i-1}\right]\right) \quad (4)$$

In Eq. (4), H is a nonlinear transformation function. i is the number of layers in network x. The dense connection structure of DenseNet can be illustrated with the help of Fig. 1.

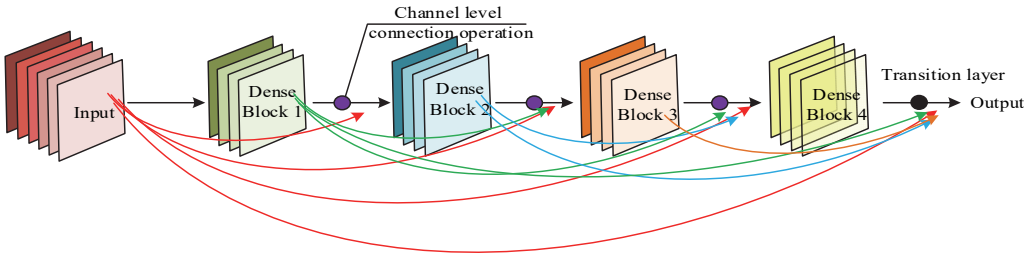


Figure 1 Schematic diagram of dense connection structure of DenseNet architecture

The core module of DenseNet is Dense Block, which achieves connections at various levels through channel level connections. The aggregation of levels leads to a linear increase in the number of channels, and the aggregation ability of the middle layer on the feature layer will affect the aggregation ability of the last layer [29]. To ensure the expressive power of the final feature map, the VoVNet model has designed an OSA module to avoid the drawbacks caused by dense connections in Dense Blocks. Fig. 2 is a diagram of the OSA module.

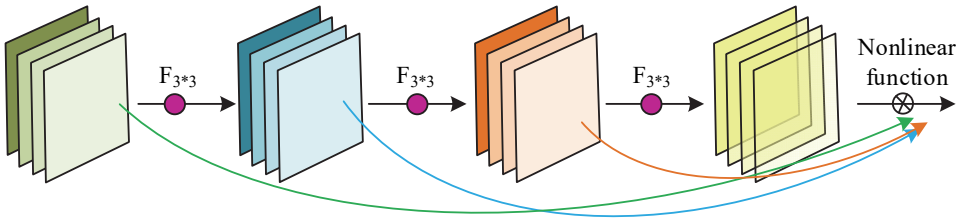


Figure 2 Schematic diagram of the OSA module

The OSA module can reduce MAC. Each level is connected by convolutional layers, which can aggregate data features while obtaining richer feature maps. The configuration structure of the OSA module includes three types, and at the end of each stage, a 3 × 3 max pooling layer with a structure of 2 is used for downsampling, resulting in an output structure of 32. The VoVNet structure, as a lightweight model, reduces the feature dimension after being processed by the OSA module. Fig. 3 shows the overall process of the VoVNet-CRF model.

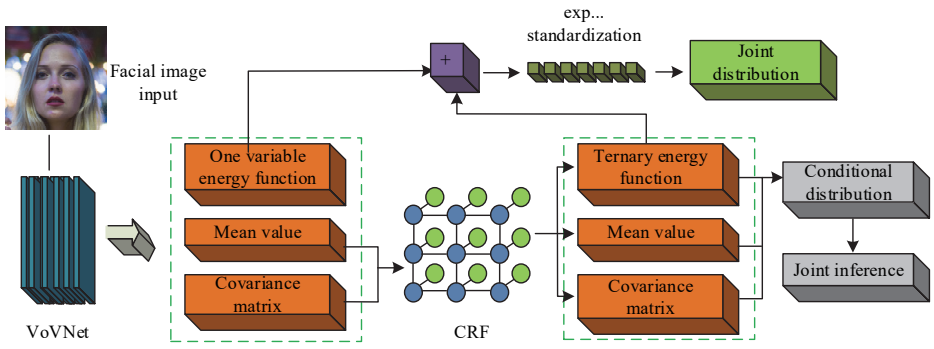


Figure 3 Schematic diagram of the overall process of VoVNet-CRF model

Common DL networks often utilize softmax or L_1 loss functions. To improve the computational and processing efficiency of the model, this study adds a negative logarithmic likelihood loss function (Gaussian NLLLoss, NLL) with true labels following a Gaussian distribution on the basis of the softmax + L_1 average loss function. The mean and variance of the sample batch data in this function are the output values of the neural network, and the output values satisfy the Gaussian distribution. Eq. (4) is the mathematical expression of NLL.

$$l(\theta) = -\sum_{i=1}^{m} \log\left(P_G\left(x^i; \theta\right)\right) \tag{4}$$

In Eq. (4), θ is a variable that adapts to random changes. x^i is the pixel point. $P_G(x^i; \theta)$ is the distribution model. The smaller the negative log likelihood value, the better the expected estimate. The softmax function can compress multidimensional vectors. The NLL function can extract the output value and corresponding label value, without logarithmic operation, only using the logarithmic value of the normalization function on the activation function.

3.2 Facial Keypoint Localization and Recognition Based on DL

In unconstrained facial keypoint localization, there is a highly nonlinear relationship between the facial appearance contour and the position of keypoints. Based on the consideration of the balance between local features and global constraints as well as robustness, facial keypoint

detection with the help of CNN-CRF is studied. Among them, CNN is good at extracting local texture features, which is suitable for pixel-level localisation, while CRF models the spatial dependency between keypoints through probabilistic graphical models, which can solve the local perception limitation of CNN. ViT requires large-scale pre-training and is not sensitive enough to local details; GNN needs to explicitly construct face topology maps, which has high computational complexity and makes it difficult for ViT and GNN to trade off accuracy-speed performance. The current CNN deep model has good feature representation performance for shape changes, and the cascading form of its regression model can improve the efficiency of facial keypoint localization to a certain extent [30]. The limitations of deep scale and regression models in terms of function approximation ability and difficulty in extracting complex features. In view of this, this study uses CNN as a benchmark to enhance data samples, fuse features at different depths, and improve the loss function. CNN can filter multiple candidate boxes, discriminate and locate facial regions and keypoints, and extract more effective features. The cascade structure is shown in Fig. 4.

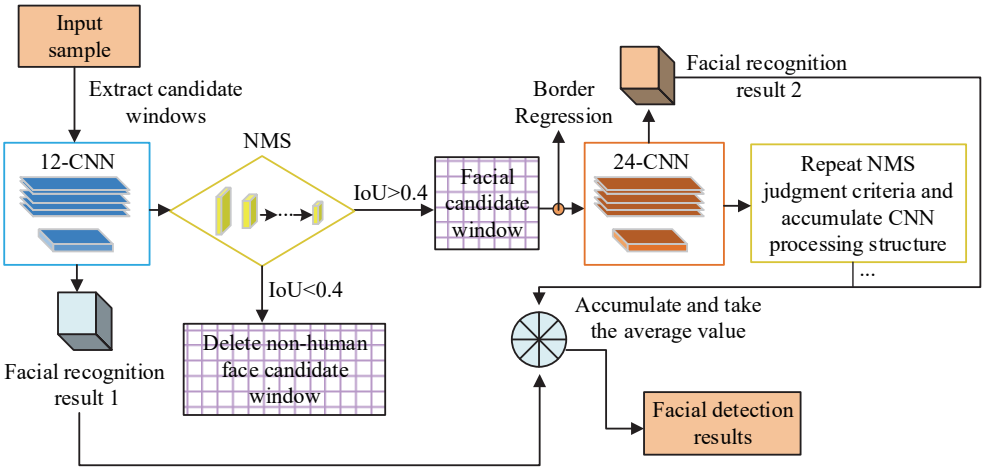


Figure 4 Cascade structure

This result can preliminarily achieve the extraction of candidate windows, the acquisition and screening of border regression vectors, face detection, and other related tasks. The input samples are first subjected to CNN to extract candidate windows and regression vectors, and non maximum suppression is used to remove irrelevant windows. The remaining windows are input into a 24 layer CNN structure for border regression calibration and filtering of the remaining windows. The input samples include both facial and non facial regions [31]. The sample is calculated using the cross entropy function to achieve classification and division of facial regions. The expression is Eq. (5).

$$L_i = -\left(y_i \log\left(p_i\right) + \left(1 - y_i\right)\left(1 - \log\left(p_i\right)\right)\right) \tag{5}$$

In Eq. (5), p_i represents the probability of a face, and y_i represents the labeled face data. When selecting candidate windows, sample partitioning is achieved by calculating the overlap of the windows, and the calculation formula is Eq. (6).

$$I_o U = S_{A \cap B} / S_{A \cup B}) \tag{6}$$

In Eq. (6), A represents the highest window. B is the testing window. $S_{A \cap B}$ is the intersection of windows. $S_{A \cup B}$ is the union of windows. $I_o U$ is the border overlap degree. If the calculated overlap value is greater than 0.65, the result is a positive sample; if it is less than 0.40, the result is a negative sample. The selected candidate window can be made closer to the real window using border regression method. The expression for calculating window offset using the Euclidean loss function is Eq. (7).

$$L_i^{\text{box}} = \left\| \hat{y}_i^{\text{box}} - y_i^{\text{box}} \right\|_2^2 \tag{7}$$

In Eq. (7), $\hat{y}_i^{\text{box}}$ is the regression target obtained by the neural network, and y_i^{box} is the label vector data of the facial block diagram.

This study uses DCNN as the benchmark network to improve its keypoint localization network structure to achieve the localization of keypoints such as eyes, nose tip, and mouth corners. This study solves the problem of feature space information loss by constructing a fast connected convolutional layer, which fuses the output features of the low and high layers of the network and connects them to a fully connected layer to ensure the acquisition of spatial contextual feature information. Each shortcut link structure has a convolutional layer that can perform convolution operations on the underlying information, increasing the network width while also enhancing feature mining capabilities. The schematic diagram of the global regression neural network structure is shown in Fig. 5.

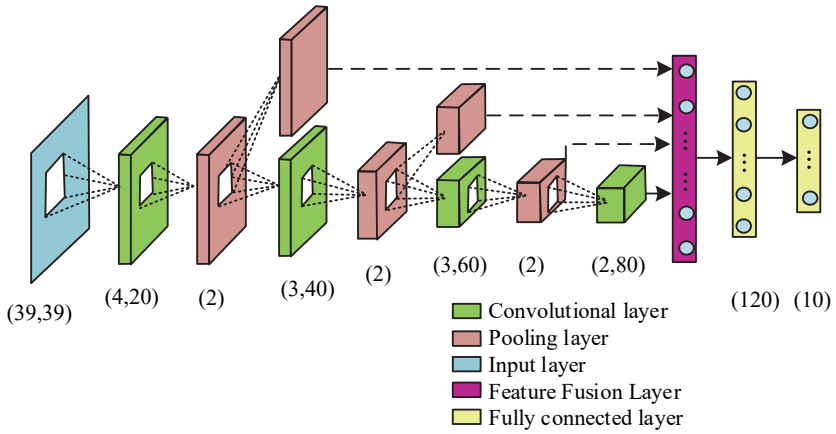


Figure 5 Global regression neural network structure

In this structure, the input layer size is 39 × 39. Each convolutional layer is followed by a corresponding activation function. The feature fusion layer can fuse the output results of the pooling layer and the last convolutional layer. The eigenvector is 10 dimensions. The L_2 norm function in the original DL may result in excessive penalty when dealing with large differences between predicted and true values, leading to biased results [32]. The cascaded regression model will locate the optimal shape increment for the objective function at different stages, and iteratively process the shape offset obtained from the regression function until the current error rate no longer decreases. The L_1 function of this model has a large deviation variation in the case of outliers, which can easily lead to gradient explosion problems, resulting in poor mapping expression ability of the model. Therefore, this study proposes a Moderate Loss function that can consider different biases and regression scenarios. It combines the advantages of L_2 and L_1, minimizing the absolute difference between the predicted value and the target value, while adding a learnable parameter to control the absolute difference. The formula for the Moderate Loss function is Eq. (8).

$$\mathrm{MLoss}(d)=\begin{cases}2d^2 & ,\ \text{if } |d|<1\\ \alpha|d| & ,\ \text{otherwise}\end{cases} \tag{8}$$

In Eq. (8), d is the absolute difference and α is the learnable parameter. This function can adjust the control parameters when dealing with facial keypoint localization problems. When DL builds a cascaded network, the first level output serves as the input for subsequent regression results. After global regression processing in the next level, the facial keypoint coordinates of the final localization result can be obtained.

3.3 Experimental Environment Condition Setting

This study utilizes the DL framework platform (Caffe), programming language (Python), MySQL open-source database, and pymysql third-party library to complete the experimental setup. The experimental hardware equipment is Windows 10 system (CPU: Intel i7-7700; Memory: 16 GB; Solid-state drive: 128 GB). The HBVCAM high-definition camera (1 million pixel 720P) is selected for the experiment. The camera includes a microphone module and an all-in-one facial recognition USB drive, which meets the requirements of facial recognition. This study utilizes a 300 W dataset. The images in this dataset cover facial images with different poses, expressions, changes in lighting conditions, and environmental scene transitions. Using Python programs to convert image sizes. The images are uniformly converted to a size of 600 × 600, and the data is randomly divided into a training set and a testing set according to an 8:2 division ratio.

4 Results and Discussion

4.1 Facial Keypoint Detection Performance

Facial keypoint detection is the core preprocessing link of face recognition system, and its positioning accuracy directly affects the accuracy of subsequent feature extraction and matching. In this section, we evaluate the detection performance of the model on facial keypoint features with the help of standard Normalised Mean Error (NME), detection accuracy and other indicators. The research algorithm is compared with DCNN, object detection algorithm (You Only Look Once Version 4 tiny, YOLOv4-tiny), and CNN-CRF algorithm. The results are shown in Fig. 6.

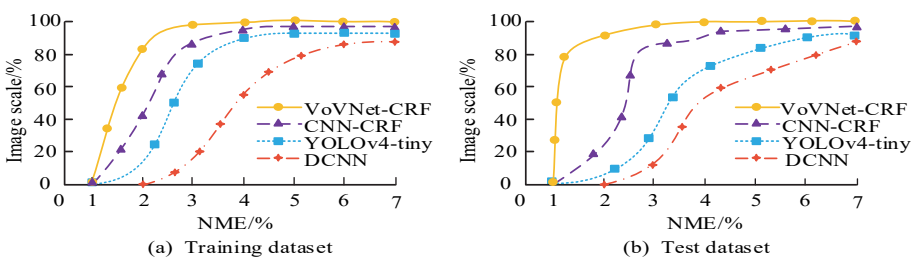


Figure 6 Standard normalized average error of different algorithms

Fig. 6 shows that on the training and testing datasets the research model exhibits relatively small cumulative error results, and the amplitude of the curve variation nodes is relatively small. The NME values of DCNN, YOLOv4-tiny, and CNN-CRF increase with the increase of graph scale, and their robustness performance is poor on different datasets. Specifically, the NME exhibited by the research method quickly converged to a value of 3% under the maximum image proportion on the training and testing datasets. However, the DCNN algorithm can only achieve 100% image scale at around 5% NME, and its performance on the test dataset is poor. The reason for this result is that the lightweight design of VoVNet in the research model reduces the feature degradation problem of deep networks. The CRF model models the geometric constraints between keypoints through probability graphs (such as the relative positions of eyes and nose), which can ensure the determination of facial keypoint position structure based on the analysis of keypoint geometric relationships. It can effectively enhance the localization ability of subtle facial features (such as corners of the eyes and mouth), thereby reducing errors. And the cascaded network structure effectively improves the problem of feature space information loss, making up for the weak adaptability of CNN-CRF to dynamic deformation. In contrast, deep convolutions of DCNN and YOLOv4 tiny may lose local details. The high error of DCNN is due to its cascaded structure being prone to losing local features under unconstrained conditions (such as occlusion) and lacking global relationship modeling; YOLOv4 tiny, as an object detection model, has insufficient adaptability to fine-grained localization of key points, resulting in a significant increase in NME with increasing image size ratio. Although CNN-CRF introduces probability maps, its feature extraction ability based on CNN is weaker than VoVNet, especially unstable under complex conditions. Subsequently, the FKPD recognition performance of the four algorithms is tested using Accuracy (ACC). ACC indicator can accurately determine the probability of consistency between test results and real results. After calculating the facial keypoint features in the dataset, Fig. 7 is obtained.

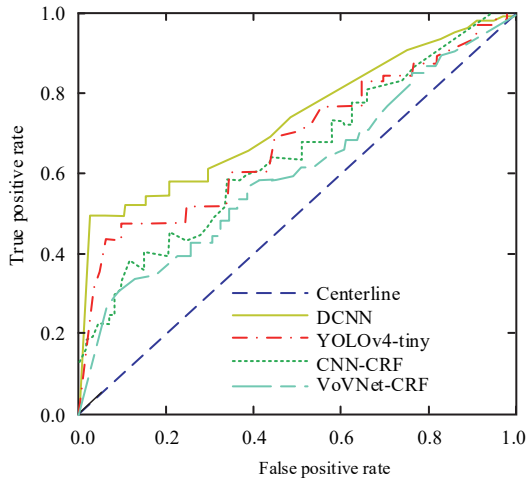
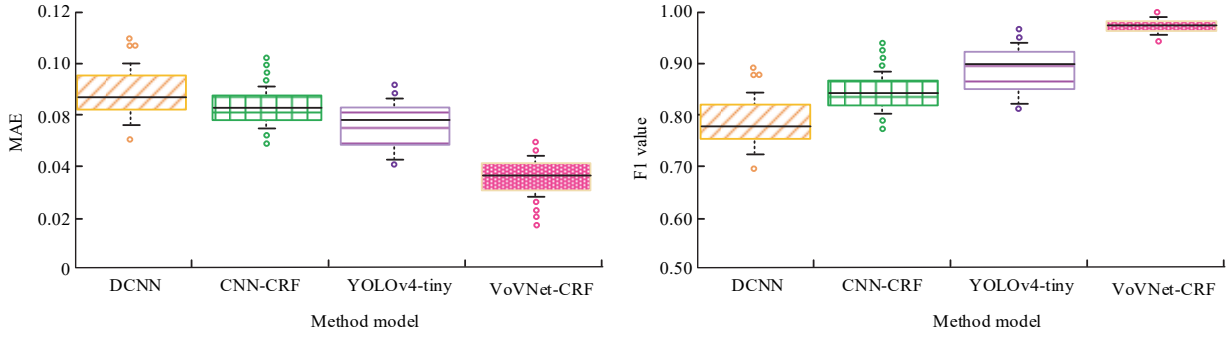


Figure 7 Accurate performance of facial keypoint feature recognition algorithm

In Fig. 7, there are differences in the overall accuracy of facial feature recognition among different algorithms. The performance ranges from good to poor as follows: VoVNet-CRF > YOLOv4-tiny > CNN-CRF > DCNN, with corresponding ACC values of 0.915, 0.907, 0.894, and 0.834. VoVNet 'cross layer shortcut connections alleviate gradient vanishing, allowing it to better preserve feature information in facial feature recognition. However, deep downsampling in DCNN causes feature map degradation, making it unable to capture subtle features. ACC is 8.1% lower than the optimal model. The ACC (0.907) of YOLOv4 tiny is close to VoVNet CRF, but its anchor box design has limited generalization to dense keypoints. It should be noted that its high ACC may be due to overfitting to simple samples. Afterwards, the study analyzed the error results of different algorithms in face detection using Mean Absolute Error (MAE) and F1 value, and the results are shown in Fig. 8.



(a) MAE c (b) F1 value

Figure 8 Facial detection errors of different algorithms

In Fig. 8a, the average MAE exhibited by the proposed algorithm is less than 0.04, which is much lower than the 0.085, 0.081, and 0.078 of the DCNN algorithm, YOLOv4 tiny algorithm, and CNN-CRF algorithm. And the F1 value of the research algorithm exceeds 0.90, with an average value of 0.096, indicating good facial detection accuracy. The YOLOv4 tiny algorithm, which performs well, has a face detection accuracy F1 value approaching 0.90, while the F1 values of DCNN and CNN-CRF algorithms are less than 0.85.

4.2 Inference Speed and Delay Power Consumption Results

Inference efficiency and energy consumption are key indicators to evaluate the practicality of face recognition systems. This section analyses the transmission frame rate and power consumption results under different algorithms by designing different test environments to better provide reference support for their deployment selection. Tab. 1 analyzes the test results of facial images in the test set under different conditions.

Table 1 Index test of four algorithms in face images with different complexity

Dataset complexity	Index	YOLOv4-tiny	CNN-CRF	DCNN	VoVNet-CRF
Simple	mAP	88.56	90.14	89.33	92.07
	FPS / img/s	81.19	84.97	66.13	88.01
	Energy Efficiency / J/img	7.2	6.4	5.2	3.0
Secondary	mAP	87.28	88.86	88.05	90.79
	FPS / img/s	66.61	73.39	51.55	83.43
	Energy Efficiency / J/img	4.4	3.6	2.4	1.9
Complex	mAP	87.16	88.87	87.98	91.92
	FPS / img/s	55.32	62.17	52.26	76.14
	Energy Efficiency / J/img	5.6	4.8	3.6	2.1

The indicators used in Tab. 1 include Mean Average Precision (mAP), Frame Per Second (FPS), and energy efficiency. The simple condition refers to a facial image that is unaffected by any factors. The moderate and complex conditions mainly involve the presence of partial facial occlusion, as well as facial images with occlusion and changes in lighting and shadows. The mAP values of the research algorithm under three conditions are 92.07, 90.79, and 91.92, which are much higher than other algorithms under the same conditions. The CNN-CRF method performed better in mAP, with mAP values of 90.14, 88.86, and 88.87 under the three conditions, respectively. The reason for the above results is that the CRF in the research method compensates for the prediction of key points in occluded areas through probabilistic inference, which has good occlusion adaptability. However, DCNN's lack of global modeling resulted in a sharp drop in mAP under moderate conditions (only 82.15). The convolutional structure of CNN-CRF is difficult to adapt to the prediction accuracy under different environmental interferences compared to the cascaded structure of the research method, so its mAP performance is slightly worse than the research method. The multi-scale feature fusion of VoVNet enhances edge perception under shadow, while the single-stage detection of YOLOv4 tiny is sensitive to light changes, so the mAP value of YOLOv4 tiny under moderate and complex conditions does not exceed 88. In terms of testing efficiency and energy efficiency indicators, there are at least 3 img/s and 2 J/img differences between the comparison algorithm and the research algorithm, with the worst performing algorithm being the DCNN algorithm. The mAP advantage of VoVNet CRF comes at a higher computational cost (FPS = 35, lower than YOLOv4 tiny's 42), but its energy efficiency (4 J/img) is still better than DCNN (6 J/img), indicating that it is more suitable for energy sensitive embedded devices. The dense connection structure of the research method enhances the robustness of the model to changes in lighting and partial occlusion, while other models experience a more significant decrease in accuracy under the same interference. The testing time and memory situation of different algorithms under three testing conditions were analyzed, and the results are shown in Tab. 2.

Table 2 Time and memory conditions under different algorithm tests

Dataset complexity	Index	YOLOv4-tiny	CNN-CRF	DCNN	VoVNet-CRF
Simple	Cosine similarity	0.999646	0.999982	0.999994	0.999997
	System test time / ms	14.26	13.15	16.89	8.26
	Memory percentage / MB	11.95	13.25	11.54	8.37
Secondary	Cosine similarity	0.999538	0.999977	0.999982	0.999998
	System test time / ms	15.26	15.25	17.13	9.26
	Memory percentage / MB	11.64	13.143	13.25	9.54
Complex	Cosine similarity	0.999429	0.999971	0.99997	0.999998
	System test time / ms	18.42	18.57	20.01	10.26
	Memory percentage / MB	11.84	14.32	12.52	9.35

In Tab. 2, the larger the value of cosine similarity, the more significant the decrease in the accuracy of model quantization. The results in Tab. 2 indicate that there is a significant variation in the performance of different algorithms on datasets of different complexity levels. Among them, the similarity of the studied algorithm under the three testing conditions is greater than 0.999995, which is higher than the similarity values of other compared algorithms. Moreover, the accuracy values of other algorithms are more easily affected by the difficulty of data information recognition. The testing time and memory usage of the research algorithm perform the best, with the lowest testing time and memory consumption of 10.26 ms and 8.37 MB, respectively, demonstrating good testing

stability. The testing time for other algorithms is greater than 10 ms, and the memory consumption of the CNN-CRF algorithm is more significant. The prediction results and false detection rate of facial keypoints are analyzed and compared with the actual output keypoint values. The initial learning rate is 0.001. The loss situation before and after the improvement of the neural network model is analyzed, and Fig. 9 is obtained.

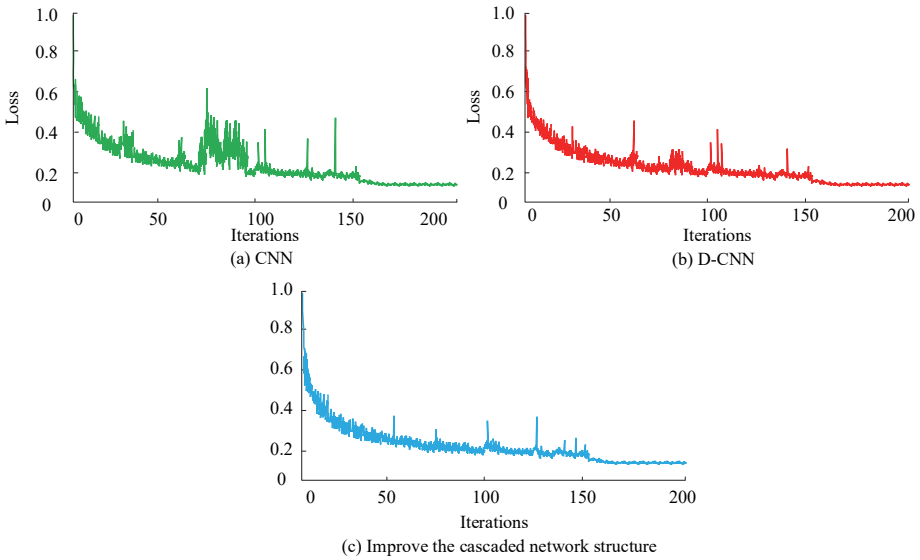


Figure 9 Loss function curve

Fig. 9a to Fig. 9c show the loss function curves of CNN, D-CNN, and research algorithms. Fig. 9 shows that the research algorithm exhibits a smaller loss function compared to other models. The fluctuation nodes of its loss function curve have less variation. The reason is that the loss function used in the study combines the advantages of other functions, which can improve data processing accuracy and reduce information loss. The improvement of the stability of the cascaded structure stems from the combination of L_1 loss (robust to outliers) and geometric constraint loss (penalty for keypoint spacing), while DCNN only uses MSE loss and is susceptible to occlusion interference.

4.3 Baseline Comparison Results

In order to further test the application performance of the proposed method of the study, it is analysed with the mainstream baseline model for locating the effect, and the ablation experiments and visualisation results are used to present the effectiveness of the research method in face recognition accuracy. The research algorithm is compared with the facial local feature recognition algorithm (SITF-LBP-Adaboost), the deep neural network fused with spatio-temporal domain features (Improved CNN), and the methods in references [19-22] for keypoint localization, as listed in Tab. 3.

Table 3 Results of evaluation indicators for facial keypoint localization

Model	NME					FR				
	Large attitude angle	Exaggerated expressions	Extreme lighting environment	Partial occlusion	Fuzzy distortion	Large attitude angle	Exaggerated expressions	Extreme lighting environment	Partial occlusion	Fuzzy distortion
Improved CNN	9.04	8.36	5.26	8.73	19.59	6.66	5.07	11.44	7.81	11.56
SITF-LBP-Adaboost	8.88	8.15	5.14	8.57	19.42	6.51	4.91	11.28	7.64	11.34
Zhang, Z. [22]	8.88	8.27	5.15	8.57	19.41	6.45	4.57	10.89	7.52	10.25
Harakannanavar, S. [19]	8.71	7.96	4.92	8.39	19.23	6.32	4.73	11.13	7.46	11.18
Zhang, F. [20]	8.37	7.62	4.61	8.06	18.86	5.99	4.34	10.77	7.13	10.73
Lv C [21]	8.44	7.63	4.67	8.09	18.89	6.02	4.37	10.82	7.16	10.78
Research model	7.39	5.51	2.61	3.08	6.88	4.01	2.36	7.79	5.15	6.76

In Tab. 3, the research model shows lower NME values compared to other models. It has good positioning accuracy under conditions of large posture angles, exaggerated expressions, extreme lighting, partial occlusion, and fuzzy distortion subdivision, with a maximum NME value not exceeding 8. Secondly, the improved local binary method and 3D facial feature method proposed by Zhang et al. and Lv et al. performed well, with minimum NME values of 4.61 and 4.67 under different conditions. However, the improved local binary mode may be limited by specific facial expressions or lighting conditions, and the 3D facial feature method is not suitable for extreme poses or facial expression changes, resulting in lower facial localization accuracy than the research model. The SITF LBP Adaboost method combines the scale invariance of SIFT and the texture description ability of LBP, resulting in an NME value of no more than 6 under extreme lighting conditions. In terms

of error rate indicators, the FR values of the comparison algorithm are above 10 in both cases of fuzzy distortion and extreme lighting environments. The difference between the improved CNN, SITF-LBP-Adaboost, literature methods in [19-22], and the research model under these two conditions exceeds 3. The reason is that the Improved CNN algorithm ignores the structural issues between facial keypoint features, and its localization performance has corresponding limitations. Although the SITF-LBP Adaboost algorithm considers the problem of data feature differences, its processing method for facial images still has the hidden problem of information loss, so the localization performance of these two algorithms is poor. It should also be noted that SITF LBP Adaboost feature extraction may be time-consuming and sensitive to noise, which may lead to an increase in model complexity and computational cost. The multi column convolutional neural network method proposed by Zhang et al. has limitations in that it may rely on high-quality depth image inputs to refine the depth information of facial image keypoints. Improving accuracy requires more complex models and more computing resources, and in practical applications, the relationship between the two needs to be balanced based on specific requirements. But overall, the research methods have shown good overall performance. The results of further analysis of the model performance using ablation experiments are displayed in Fig. 10.

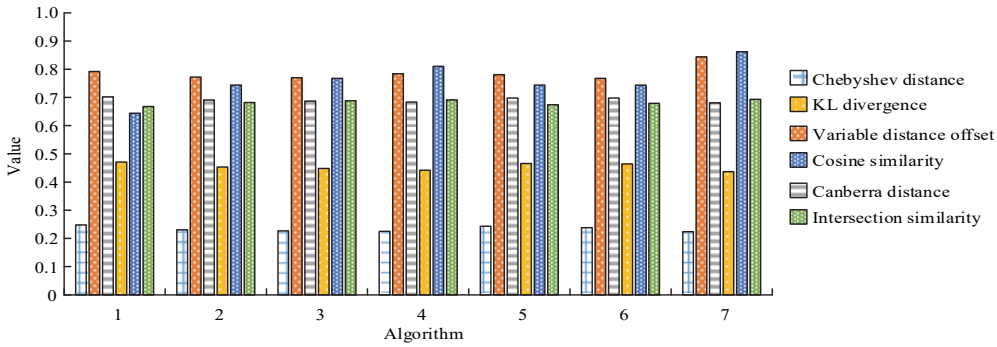


Figure 10 Algorithm ablation experimental results

In Fig. 10, the Chebyshev distance value of the research algorithm is 0.223, significantly lower than other algorithms, and its numbers in cosine coefficient and intersection similarity indicators are even lower than 0.9, indicating a high degree of similarity in numerical calculations. This indicates that it has a significant effect on extracting key point features and can achieve good position localization. This study conducts a visual analysis of the proposed localization model, as shown in Fig. 11.

Figure 11 Localization effect of facial keypoints

Tests have shown that the research model has good keypoint localization results for faces with different posture angles, high accuracy in locating facial feature position information, and good robustness.

4.4 Discussion

The smaller the normalized mean error (NME) value, the better the robustness of the algorithm. The results of facial keypoint detection indicate that the NME value of the proposed keypoint detection model is much lower than other algorithms, and it has good robustness on different datasets. Experiments have shown that the VoVNet-CRF model (0.915) outperformed YOLOv4-tiny (0.907), CNN-CRF (0.894), and DCNN (0.834) in overall facial feature recognition accuracy. This is because combining the VoFHIR model with CRF can better represent and process features. Aggregation strategy can expand the

receptive field. The CRF fully connected form can combine and adjust the classification results of deep learning with the original image elements to clarify edges. Therefore, the model has good adaptability when dealing with different datasets. Although the DCNN algorithm can gradually extract higher-level features through multiple convolutional layers, its fixed receptive field makes it difficult to capture all key points, and as the image scale increases, detail information may be lost. The lightweight design of YOLOv4 micro algorithm limits its feature extraction ability, while the robustness of CNN-CRF algorithm depends on the selection of its features and parameters, resulting in unstable feature optimization results during global processing. Unlike the convolutional neural network architecture proposed by Wen et al. [10] for information recognition, the method proposed in this study can determine the position structure of facial keypoints while ensuring the analysis of their geometric relationships, and the recognition accuracy is slightly better than the method in this literature. However, the method proposed by Wen et al. [10] is difficult to adapt to various testing scenarios. The method of combining deep convolutional neural networks and multilayer perceptrons proposed by Hariri et al. [12] has similarities with the research method, but multilayer perceptrons are difficult to control overfitting problems. The probability graph model and shortcut connection convolutional layer designed for research can effectively organize and analyze feature data, improving algorithm accuracy and efficiency while reducing losses.

The accuracy of facial keypoint feature recognition and image testing results also indicates that this method has good recognition accuracy, predictive performance, and low testing time and loss function. The curve fluctuation of the research algorithm in facial keypoint feature recognition was relatively small, with mAP values of 92.07, 90.79, and 91.92 under three image conditions, which were much higher than other algorithms under the same conditions. The reason is that the adaptive advantage of CRF can effectively ensure the image processing effect of the algorithm under different conditions, reducing the accuracy degradation caused by occlusion. The loss function used in this study combines the advantages of other functions, which can effectively improve data processing accuracy and reduce information loss. The model proposed in this study has good localization accuracy under conditions of large pose angles, exaggerated expressions, extreme lighting, partial occlusion, and fuzzy distortion subdivision, with a much lower error rate than other algorithms. Unlike the improved local binary pattern approach proposed by Yan et al. [18], the proposed method has good adaptability and its feature extraction accuracy is also due to the algorithm. The difference between the improved CNN and SITF LBP Adaboost algorithms and the research model exceeds 3. The reason is that the improved CNN algorithm ignores the structural issues between facial keypoint features, and its localization performance also has corresponding limitations. Although the SITF-LBP Adaboost algorithm considers the issue of data feature differences, there is still a risk of information loss in its facial image processing method, resulting in poor localization performance of these two algorithms. However, Yu et al. [23] proposed using Blaze_ghost network to achieve facial keypoint recognition. Although this method can effectively improve the normal mean error, it is difficult to consider resource waste and computation time issues compared to research methods. In summary, the facial recognition model proposed by the research institute has shown good generalization performance and robustness in keypoint localization, which can provide reference value for the field of facial recognition.

YOLOv4 tiny has the highest FPS and is suitable for real-time detection on mobile devices, but its accuracy (ACC = 0.907) and complex condition mAP (decreased by about 10%) are significantly lower than the method proposed in the study. The energy efficiency of the method proposed by the research institute is better than DCNN, which is suitable for embedded devices. The experiment shows that the proposed VoVNet CRF model combines the lightweight features of VoVNet with the spatial probability modeling of CRF, overcoming the limitations of CNN fixed receptive field (such as DCNN) and mixed method parameter sensitivity (such as CNN-CRF), and reducing NME by 40% compared to BlazeGhost and others under occlusion and extreme poses. The model performs stably on different datasets (mAP > 90% under complex conditions) and has excellent energy efficiency (4 J/img), making it suitable for monitoring and AR scenarios. But there is a trade-off: the CRF layer increases inference time by 15% and relies on high-quality training data, limiting its scalability in low resolution images. Further optimization is needed in the future.

5 CONCLUSION

The experimental results and comparative analysis jointly indicate that the proposed VovDCRF model achieves a significant balance between accuracy, robustness, and computational efficiency in facial keypoint localization. The improved cascaded structure algorithm exhibited a smaller loss function when processing facial keypoint data. Compared with other models, the improved model had better positioning accuracy under conditions of large pose angles, exaggerated expressions, extreme lighting, partial occlusion, and fuzzy distortion subdivision. The ablation experiment showed that the research algorithm had a significant effect on extracting keypoint features, and it had good keypoint localization accuracy for faces with different poses and angles. In summary, the model proposed by the research institute has demonstrated excellent accuracy and efficiency in facial keypoint detection under various conditions. Its geometric preservation feature aggregation and occlusion adaptive inference make it highly adaptable. However, the real-time performance of this model indicates that hardware aware optimization, such as pruning or quantization, is needed to meet the needs of edge computing. Future work will focus on embedding attention mechanisms for dynamic feature refinement and integrating meta learning to reduce data dependencies and broaden their applicability. Strengthening the efficiency of model processing and mapping low-level visual features, and combining them with reinforcement learning and transfer learning methods to achieve cross domain generalization and self supervised

learning technology integration, is an important focus of future research.

6 REFERENCE

[1] Li, S. & Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, *13*(3), 1195-1215. https://doi.org/10.1109/TAFFC.2020.2981446

[2] Andrejevic, M. & Selwyn, N. (2020). Facial recognition technology in schools: Critical questions and concerns. *Learning, Media and Technology*, *45*(2), 115-128. https://doi.org/10.1080/17439884.2020.1686014

[3] Wang, Z., Huang, B., Wang, G., et al. (2023). Masked face recognition dataset and application. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *5*(2), 298-304. https://doi.org/10.1109/TBIOM.2023.3242085

[4] Dubey, S. & Meena, R. (2021). A review of face recognition using SIFT feature extraction. *IOSR Journal of VLSI Signal Processing*, *5*(2), 31-35. https://doi.org/10.1109/TBIOM.2023.3242085

[5] Zhu, Z., Huang, G., Deng, J., Ye, Y., Huang, J., Chen, X., Zhu, J., Yang, T., Du, D., Lu, J., & Zhou, J. (2022). Webface260M: A benchmark for million-scale deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(2), 2627-2644. https://doi.org/10.1109/TPAMI.2022.3169734

[6] Lin, W., He, X., Dai, W., et al. (2020). Key-point sequence lossless compression for intelligent video analysis. *IEEE MultiMedia*, *27*(3), 12-22. https://doi.org/10.1109/MMUL.2020.2990863

[7] Alkhazali, S. & El-Bashir, M. (2020). Local binary pattern method (LBP) and principal component analysis (PCA) for periocular recognition. *International Journal of Advanced Computer Science and Applications*, *11*(8), 1-7. https://doi.org/10.14569/IJACSA.2020.0110810

[8] Yu, Z., Qin, Y., Li, X., et al. (2022). Deep learning for face anti-spoofing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(5), 5609-5631. https://doi.org/10.1109/TPAMI.2022.3215850

[9] Bisogni, C., Nappi, M., Pero, C., et al. (2021). PIFS scheme for head pose estimation aimed at faster face recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *4*(2), 173-184. https://doi.org/10.1109/TBIOM.2021.3122307

[10] Wen, T., Ding, Z., Yao, Y., et al. (2022). Picassonet: Searching adaptive architecture for efficient facial landmark localization. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(12), 10516-10527. https://doi.org/10.1109/TNNLS.2022.3167743

[11] Gupta, S., Thakur, K., & Kumar, M. (2021). 2D-human face recognition using SIFT and SURF descriptors of face's feature regions. *The Visual Computer*, *37*(3), 447-456. https://doi.org/10.1007/s00371-020-01814-8

[12] Ahila Priyadharshini, R., Arivazhagan, S., & Arun, M. (2021). A deep learning approach for person identification using ear biometrics. *Applied Intelligence*, *51*(4), 2161-2172. https://doi.org/10.1007/s10489-020-01995-8

[13] Vu, H. N., Nguyen, M. H., & Pham, C. (2022). Masked face recognition with convolutional neural networks and local binary patterns. *Applied Intelligence*, *52*(5), 5497-5512. https://doi.org/10.1007/s10489-021-02728-1

[14] Hariri, W. (2022). Efficient masked face recognition method during the COVID-19 pandemic. *Signal, Image and Video Processing*, *16*(3), 605-612. https://doi.org/10.1007/s11760-021-02050-w

[15] Misael Burruel-Zazueta, J. M., Rodríguez-Rangel, H., Peralta-Peñuñuri, G. E., et al. (2024). Biometric lock with facial recognition implemented with deep learning techniques. *Computer Science and Information Systems*, *21*(4), 1359-1387. https://doi.org/10.2298/CSIS240229038B

[16] Niu, M., Li, M., & Fu, C. (2024). Point Transform Networks for automatic depression level prediction via facial keypoints. *Knowledge-Based Systems*, *297*, 111951. https://doi.org/10.1016/j.knosys.2024.111951

[17] Li, S. & Deng, W. (2020). A deeper look at facial expression dataset bias. *IEEE Transactions on Affective Computing*, *13*(2), 881-893. https://doi.org/10.1109/TAFFC.2020.2973158

[18] Cowen, A. S., Keltner, D., Schroff, F., et al. (2021). Sixteen facial expressions occur in similar contexts worldwide. *Nature*, *589*(7841), 251-257. https://doi.org/10.1038/s41586-020-3037-7

[19] Harakannanavar, S., Sapnakumari, C., Ramachandra, A., et al. (2023). Performance evaluation of fusion based efficient algorithm for facial expression recognition. *Indian Journal of Science and Technology*, *16*, 266-276. https://doi.org/10.17485/IJST/v16i4.1891

[20] Zhang, F., Zhang, T., Mao, Q., et al. (2020). Geometry guided pose-invariant facial expression recognition. *IEEE Transactions on Image Processing*, *29*, 4445-4460. https://doi.org/10.1109/TIP.2020.2972114

[21] Lv, C., Wu, Z., Wang, X., et al. (2020). 3D facial similarity measurement and its application in facial organization. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *16*(3), 1-20. https://doi.org/10.1145/3397765

[22] Zhang, Z., Lian, D., & Gao, S. (2021). RGB-D-based gaze point estimation via multi-column CNNs and facial landmarks global optimization. *The Visual Computer*, *37*(7), 1731-1741. https://doi.org/10.1007/s00371-020-01934-1

[23] Liu, P., Lin, Y., Meng, Z., et al. (2021). Point adversarial self-mining: A simple method for facial expression recognition. *IEEE Transactions on Cybernetics*, *52*(12), 12649-12660. https://doi.org/10.1109/TCYB.2021.3085744

[24] Dar, A. S. & Palanivel, S. (2022). Real time face authentication system using stacked deep auto encoder for facial reconstruction. *International Journal of Thin Film Science and Technology*, *11*(1), 73-82. https://doi.org/10.18576/ijtfst/110109

[25] Wu, X., Ma, Y., Lian, H., et al. (2023). Clustering optimized portrait matting algorithm based on improved sparrow algorithm. *Tehnički vjesnik - Technical Gazette*, *30*(6), 1911-1919. https://doi.org/10.17559/TV-20230701000778

[26] Minaee, S., Abdolrashidi, A., Su, H., et al. (2023). Biometrics recognition using deep learning: A survey. *Artificial Intelligence Review*, *56*(8), 8647-8695. https://doi.org/10.1007/s10462-022-10237-x

[27] Revina, I. M. & Emmanuel, W. R. S. (2021). A survey on human face expression recognition techniques. *Journal of King Saud University - Computer and Information Sciences*, *33*(6), 619-628. https://doi.org/10.1016/j.jksuci.2018.09.002

[28] Yang, Y. & Song, X. (2022). Research on face intelligent perception technology integrating deep learning under different illumination intensities. *Journal of Computational and Cognitive Engineering*, *1*(1), 32-36. https://doi.org/10.47852/bonviewJCCE19919

[29] Yaagoup, K. M. M. & Mus, M. E. M. (2020). Online Arabic handwriting characters recognition using deep learning. *International Journal of Advanced Research in Computer and Communication Engineering*, *9*(10), 83-92. https://doi.org/10.17148/IJARCCE.2020.91014

[30] Zhang, J., Yu, X., Lei, X., et al. (2022). A novel deep LeNet-5 convolutional neural network model for image recognition. *Computer Science and Information Systems*, *19*(3), 1463-1480. https://doi.org/10.2298/CSIS220120036Z

[31] Wang, H. & Hou, S. (2020). Facial expression recognition based on the fusion of CNN and SIFT features. *2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, 190-194. https://doi.org/10.1109/ICEIEC49280.2020.9152361

[32] Gao, P., Lu, K., Xue, J., et al. (2021). A facial landmark detection method based on deep knowledge transfer. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(3), 1342-1353. https://doi.org/10.1109/TNNLS.2021.3105247

Contact information:

Li Juan YANG
(Corresponding author)
North China Institute of Aerospace Engineering,
Hebei, China, 065000
E-mail: lijuanyang168@163.com

Ying LI
North China Institute of Aerospace Engineering,
Hebei, China, 065000
E-mail: xue113688@163.com