Computers, Materials & Continua
DOI: 10.32604/cmc.2022.027767
Article

Tech Science Press

Research on Facial Expression Capture Based on Two-Stage Neural Network

Zhenzhou Wang[1], Shao Cui[1], Xiang Wang[1,*] and JiaFeng Tian[2]

[1]School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, 050000, China
[2]School of Engineering, Newcastle University, Newcastle Upon Tyne, NE98, United Kingdom
*Corresponding Author: Xiang Wang. Email: wangxiang@hebust.edu.cn
Received: 25 January 2022; Accepted: 08 March 2022

Abstract: To generate realistic three-dimensional animation of virtual character, capturing real facial expression is the primary task. Due to diverse facial expressions and complex background, facial landmarks recognized by existing strategies have the problem of deviations and low accuracy. Therefore, a method for facial expression capture based on two-stage neural network is proposed in this paper which takes advantage of improved multi-task cascaded convolutional networks (MTCNN) and high-resolution network. Firstly, the convolution operation of traditional MTCNN is improved. The face information in the input image is quickly filtered by feature fusion in the first stage and Octave Convolution instead of the original ones is introduced into in the second stage to enhance the feature extraction ability of the network, which further rejects a large number of false candidates. The model outputs more accurate facial candidate windows for better landmarks recognition and locates the faces. Then the images cropped after face detection are input into high-resolution network. Multi-scale feature fusion is realized by parallel connection of multi-resolution streams, and rich high-resolution heatmaps of facial landmarks are obtained. Finally, the changes of facial landmarks recognized are tracked in real-time. The expression parameters are extracted and transmitted to Unity3D engine to drive the virtual character's face, which can realize facial expression synchronous animation. Extensive experimental results obtained on the WFLW database demonstrate the superiority of the proposed method in terms of accuracy and robustness, especially for diverse expressions and complex background. The method can accurately capture facial expression and generate three-dimensional animation effects, making online entertainment and social interaction more immersive in shared virtual space.

Keywords: Facial expression capture; facial landmarks; multi-task cascaded convolutional networks; high-resolution network; animation generation

1 Introduction

Facial expression capture technology is a major research in computer graphics character animation and other fields. With the emergence of shared virtual space, synchronous avatar facial expression can convey human emotions and effective social experience, which plays a crucial role in digital entertainment, film and television animation industry and other fields [1]. For example, movies like *Avatar*, which is well known to all, use 3D facial expression capture and generation technology by computer to make the virtual character expression animation more realistic [2]. Facial expression features are extracted based on artificial intelligence technology, so that the information endows the virtual character with authentic facial expressions and emotions. The interaction between users and information will contribute to the development of the metaverse, which is an important evolution of future social ways [3,4].

Extracting effective landmarks from images is the core step of facial expression capture and the obtained landmarks can be used as the basic data of facial expression parameters [5]. Recent studies [6,7] have shown that the technique is limited by expensive capture tools, complex model reconstruction of facial expression and the inaccuracy of extracting facial landmarks using traditional image processing. It is impossible to ensure the accuracy of facial landmarks because the facial features of human face have different scale types and make different degrees of expressions. In view of these problems, this paper focuses on how to improve the accuracy of facial landmarks based on deep learning technology about diverse facial expressions and complex background, and achieve high fidelity and real-time virtual character expression animation.

Therefore, this paper proposes MTHR-Face model based on two-stage neural network for facial expression capture. Firstly, the real-time video is captured by the camera, and the target face is detected and aligned by the improved multi-task cascaded convolutional networks (MTCNN), so that a more accurate facial candidate window is obtained for landmarks recognition. Then combined with high-resolution network (HRNetV2-W18), the method realizes recognition of high resolution facial expression features and real-time tracking under different levels of expressions. Finally, the action units (AU) are extracted by regression of the recognized facial landmarks to capture facial expression, and the mapping relationship between AU and Blendshape expression bases is built to synchronize expression of the virtual character and generate animation. The overall process is shown in Fig. 1.

In summary, the contributions of this paper are as follows:

1) An improved MTCNN model is proposed to efficiently extract face feature information from images and accurately detect faces. This model can eliminate complex background and align faces, thus further reducing the difficulty of facial landmarks recognition task.
2) A two-stage neural network that improved MTCNN and HRNetV2-W18 is introduced to recognize 98 rich facial landmarks and capture expression changes. This method not only maintains the high resolution of the facial feature images in the whole process, but also makes more accurate landmarks recognition under different degrees of expressions.
3) Experiments show that the performance of MTHR-Face method on WFLW database is significantly better than other existing methods, especially for faces with difficult scenarios such as expression, large pose, and occlusion. Additionally, the algorithm can track facial expression in real-time and generate virtual character expression animation.

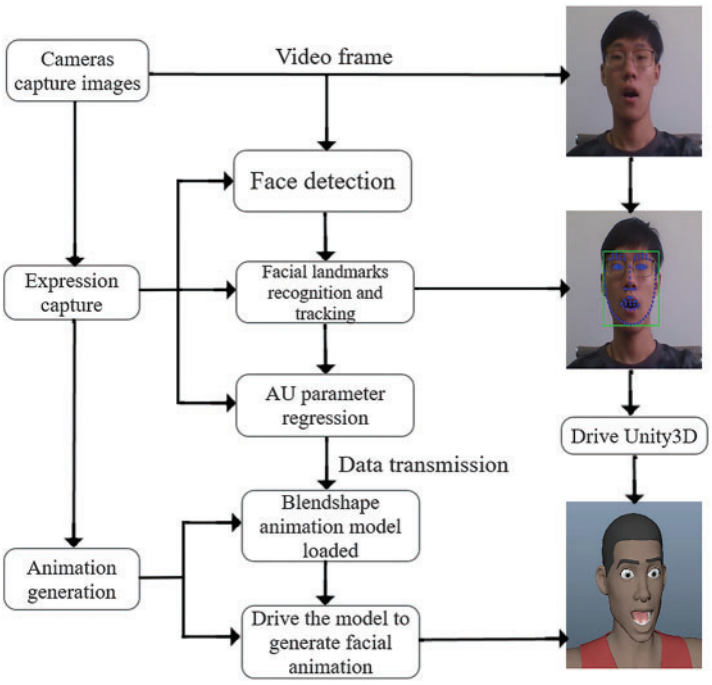

Figure 1: General flow chart of expression capture system

2 Related Work

Facial expression capture system. Sibbing et al. [8] used 5 synchronous cameras to shoot faces without markers, mainly employing the method of 3D face reconstruction to establish the connection between frames by 2D grid tracking to achieve facial expression animation reproduction, but due to the reconstruction of facial expression, it would consume more time. Cao et al. [9,10], based on constructed Face Warehouse, 3D expression library, used the regression method of extracting 3D facial expression from 2D videos to locate facial landmarks and track facial expression in real time, and generated 3D faces through registration and realistic facial expression animation with BlendShape. However, in the offline training stage, it needed to collect data from each face and generated the relatively rough facial expression model. Weise et al. [11] used Kinect motion-sensing peripherals as an expression capture tool, but the noise and error of the obtained 3D information were large. Therefore, the facial expression model was trained to reduce the noise and error to ensure the real-time animation, but the generated expression lacked variability and freedom. In recent years, facial expression capture based on deep learning has become a hot research topic. Laine et al. [12] trained the convolutional neural network based on videos of facial expressions to generate 3D expression performance by using multi-view stereoscopic tracking and enhanced animation. However, the captured facial expression data is difficult to reproduce on large multi-users and requires a lot of training data.

Facial landmarks recognition. Active Shape Model (ASM) [13], a classical algorithm in traditional methods, and AAM [14], an improved algorithm based on ASM, were used to detect facial landmarks by shape change model. The detection results of the above studies are strongly dependent on the dataset of their models, and their generalization performance and robustness are poor. Feng et al. [15] proposed a new cascade shape regression (CSR) structure for robust facial landmarks detection on unconstrained human faces. However, due to the limited capability of handmade feature

representation, there is still a problem of misalignment in facial landmarks detection using traditional methods. In recent years, the method of deep learning has greatly improved the accuracy of facial landmarks detection. After the cascade convolutional neural network (DCNN) proposed by Sun et al. [16], it was improved [17] and applied to 68 facial landmarks detection. NeWell et al. [18] used a heatmap regression model and designed a stacked hourglass network for human pose estimation. Wu et al. [19] designed the LAB boundary perception algorithm, which used eight stacked hourglass networks to predict the boundary heatmaps of facial features and decode coordinate information. Yang et al. [20] used supervised transformation of standardized faces and stacked hourglass network to obtain predictive heatmaps, which achieved good results. However, the hourglass model consumed huge computational resources and still lacked the robustness of facial landmarks detection in large-angle posture scenarios.

MTHR-Face method based on two-stage neural network leverages the best advantages of the improved MTCNN and HRNetV2-W18 model. The improved MTCNN can quickly detect faces and obtain accurate facial candidate windows, laying a foundation for facial landmarks recognition. Combined with HRNetV2-W18 model, this method not only maintains high resolution of the facial feature images in the whole process, but also improves the error of landmarks recognition under various expressions Real-time facial expression capture can realize the synchronization of virtual characters' expressions and generate animation with high fidelity. It is of great significance to real-time emotional communication and interactive control between users and avatars in shared virtual space.

3 The Traditional MTCNN Model

At present, in the field of machine vision, multi-task cascaded convolutional network (MTCNN) [21], as shown in Fig. 2, is a model of face detection and facial landmarks localization with high computational speed and good performance. It is a cascaded convolutional network architecture and composed of three-layer convolutional neural networks including Proposal Network (P-Net), Refine Network(R-Net) and Output Network (O-Net). Its main tasks are face classification, boundary box regression and facial landmarks localization respectively. Face detection and alignment as well as facial landmarks extraction are realized in the process from simplicity to refinement.

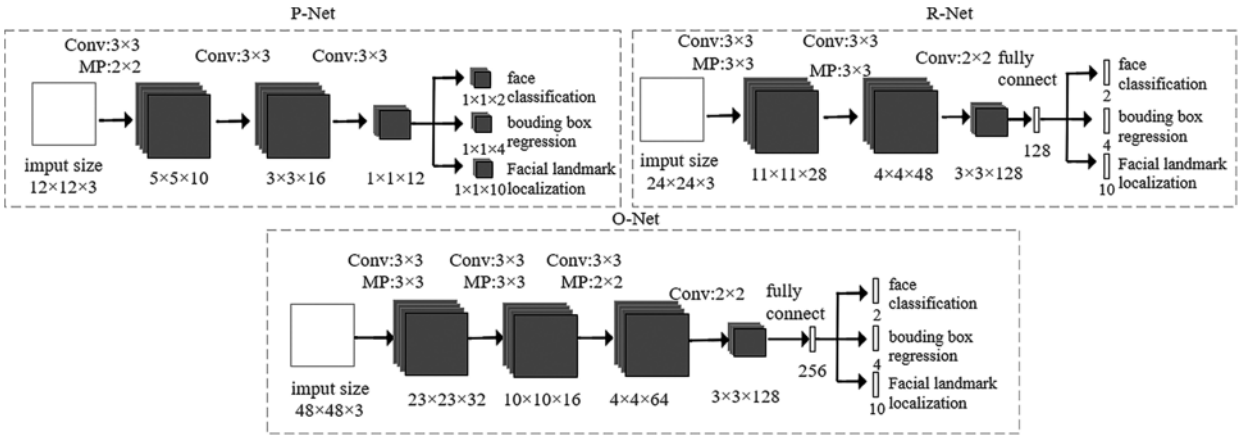

Figure 2: Structure of three-stage convolutional networks in MTCNN

This is the input of the following three-stage cascaded framework:

Step1: The operation of preprocessing ensures that the given image can be scaled to different scales to build an image pyramid, which is provided for training of the cascaded networks.

Step2: P-Net structure is a fully convolutional neural network. Images with size of 12×12 after image pyramid processing are obtained as the input of the network. In order to ensure higher recall rate, a 3×3 convolution kernel is used to carry out preliminary feature extraction through convolution operation to obtain facial candidate windows and their bounding box regression vectors. Then using the method of bounding box regression adjusts windows and non-maximum suppression (NMS) are performed to filter highly overlapped candidates. The results with bounding size of 24×24 are mapped to the original image.

A two-class classification problem of P-Net is the determination of face/non-face, so the loss function of face detection classification is cross-entropy function. The formula is as follows:

$$\mathrm{L}_{\mathrm{i}}^{\mathrm{det}} = -(y_i^{det} log(p_i) + (1 - y_i^{det})(1 - log(p_i)) \tag{1}$$

Where p_i represents the probability of the appearance of the face region, the notation $y_i^{det} \in \{0, 1\}$ represents the ground-truth label.

Step3: R-Net is basically constructed as convolutional neural network. The network will filter out a large number of poor candidate windows to achieve the effect of high-precision filtering and face region optimization, because there are many candidate windows left through P-Net. R-Net adds full connection layer to classify feature images output by P-Net, and then using fine-tuning candidate windows of boundary box vector, border regression and NMS finally are performed to remove overlapping windows for selected candidates.

R-Net is a bounding box regression problem, which is used to select the candidate windows of P-Net and filter a large number of non-face candidate windows. The Euclidean loss for each sample x_i is employed to measure the distance between the predicted and the actual value of face candidate windows. Its loss function is:

$$\mathrm{L}_{\mathrm{i}}^{\mathrm{box}} = \left\| \hat{y}_i^{box} - y_i^{box} \right\|_2^2 \tag{2}$$

where $\hat{y}_i^{box}$ indicates the regression target obtained from the network, y_i^{box} indicates the ground-truth coordinate. $y_i^{box} \in R^4$ denotes the candidate frame parameter is a vector of length 4 including left top, height and width.

Step4: The basic structure of O-Net is a relatively complex convolutional neural network. This network has more input features and in this stage will identify face regions with more supervision, further refine the coordinates of face detection windows to make the processing results more precise and meticulous. Finally, this network will output the face candidate windows and coordinates of five landmarks.

Similar to the boundary regression process, the loss function of this step is still to calculate the deviation between the predicted and the actual landmarks' position. The minimized the Euclidean loss in the process of landmarks localization is as follows:

$$\mathrm{L}_{\mathrm{i}}^{\mathrm{landmark}} = \left\| \hat{y}_i^{landmark} - y_i^{landmark} \right\|_2^2 \tag{3}$$

In the formula, $\hat{y}_i^{landmark}$ indicates the predicted coordinates, and $y_i^{landmark}$ indicates the ground-truth coordinate for the ith sample.

4 MTHR-Face Method

The primary task of facial expression capture is to accurately recognize facial landmarks. Due to the variety of facial landmarks caused by human emotion, it is impossible to guarantee the uniformity of animation generated by facial expression of virtual characters. Therefore, MTHR-Face method in this paper is adopted for facial expression capture based on improved MTCNN and HRNetV2-W18, as shown in Fig. 3. Face detection rate has a certain impact on facial features capture, which lays a foundation for the subsequent recognition of landmarks accurately, so the traditional MTCNN convolution operation is improved. Firstly, feature fusion operation is carried out in the P-Net and Octave Convolution is introduced to replace ordinary convolution in the R-Net, which can obtain more precise face candidate windows and detect target faces quickly. Then combined with HRNetV2-W18, the images cropped after face detection are input to four consecutive multi-scale cascade parallel convolution neural networks for feature fusion and obtain 98 rich high-resolution facial landmarks. The two-stage neural network can effectively maintain the spatial information and high resolution of facial feature images and make the landmarks converge under the complex geometric changes of facial expression.

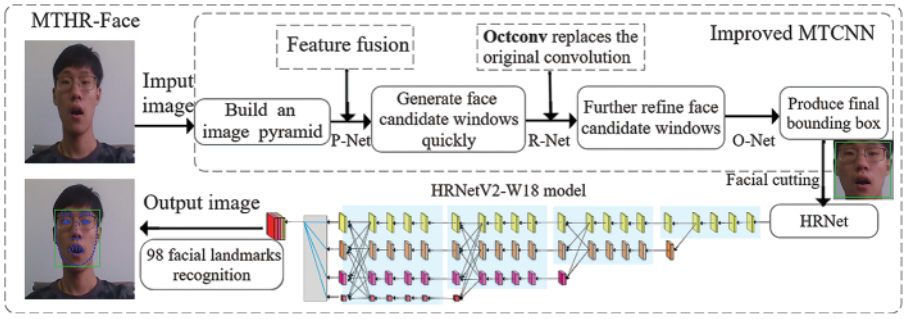

Figure 3: Block diagram of MTHR-face algorithm

4.1 Improved MTCNN for Face Detection

The purpose of face detection is to obtain appropriate facial regions to simplify the difficulty of landmarks recognition. However, efficient and accurate detection of face region has certain impact on subsequent recognition task. Therefore, traditional MTCNN convolution operation is improved. Its function is to efficiently filter the face information in the image, optimize the network structure of the model, improve the feature representation ability of the convolution layer, and ensure accurate and efficient recognition of subsequent tasks.

4.1.1 Feature Fusion Operation

P-Net as the first stage will generate a large number of overlapping face candidate windows, resulting in a long operation time. Therefore, the network is improved to shorten time, as shown in Fig. 4 below. Firstly, the input P-Net image is subjected to three-layer convolution operation and the loss of the image is calculated. The structure ① in Fig. 4 shows that multi-scale convolution operation is performed on the input image features, and then aggregation operation is performed to enhance the detection effect. Secondly, the first and second layer feature images are fused to improve the expression ability of the network and efficiently filter the face information in the image. Finally, the structure ② in Fig. 4 is the decomposition operation of the convolution kernel to reduce number of parameters and improve the detection rate of the network.

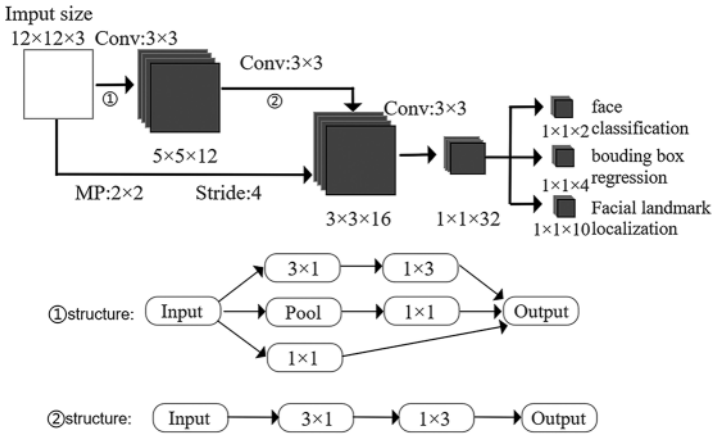

Figure 4: Improved P-Net structure

4.1.2 Introducing New Octave Convolution

Due to the small number of convolutional layers at all levels of MTCNN structure, the extracted features are not enough to fully represent face details. The whole network make full use of the features extracted from the convolution layer and increase the face detection capability of the multi-pose model, laying a foundation for facial landmarks recognition. Therefore, the new convolution operation, Octave Convolution (Octconv) [22], is introduced into R-Net of MTCNN, which only replaces the original Convolution, so that the network better realize the refinement and regression of face candidate windows and improve the detection accuracy.

The principle of Octconv is to effectively process the low frequency and high frequency components in the corresponding frequency tensor for the convolution layer to form new output features together. The specific operating principle of tensor decomposition is shown in Fig. 5 where X and Y are tensors of the input and output respectively. Output of each layer consists of low frequency and high frequency signals, and each part is composed of high frequency and low frequency components of the network output of the previous layer in a certain proportion. The formula of output high frequency signal and low frequency signal is as follows:

$$\mathrm{Y^H} = Y^{H\to H} + Y^{L\to H} \tag{4}$$

$$\mathrm{Y^L} = Y^{L\to L} + Y^{H\to L} \tag{5}$$

Fig. 5 shows that two green arrows correspond to information update of the high and low frequency characteristic graph while two red arrows facilitate information exchange between the two frequencies. The introduction of convolution can reduce the operation time and improve the face detection rate by halving the low frequency information in the input data. Compared with the features extracted by original convolution, the spatial redundancy of features is reduced and feature representation capability of convolution layer is improved.

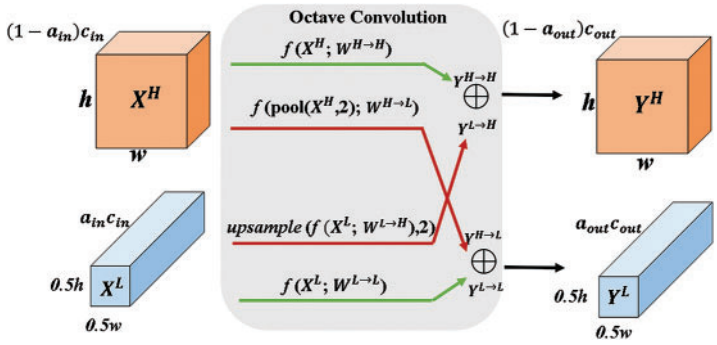

Figure 5: Detailed design of the octave convolution

4.2 Facial Landmarks Recognition Combined with High-Resolution Network

Facial landmarks recognition aims to detect the position of eyebrow, eyes, nose, mouth and facial contour from the face image. Due to the risk of missing feature information under different levels of facial expressions, more reliable landmarks are needed to further accurately capture facial expression and generate animation. Therefore, HRNetV2-W18 model [23] in this paper is selected as the face feature extractor, and the high resolution of facial feature map is maintained during the whole process, so that the prediction of facial landmarks is more accurate in space. This network with high-resolution performance is transferred to a task of facial landmarks recognition. In addition, it has achieved a good level in the MPII Human Pose database [24].

HRNetV2-W18 has four stages and is connected in parallel, as shown in Fig. 6. The first stage involves high-resolution convolution. At the beginning of each stage, a parallel stream with a lower resolution than the current minimum is added to connect the multi-resolution stream in parallel. The parallel connections of streams with different resolutions can effectively preserve the original features and extract the deep features downward. At the end of each stage, multi-scale feature fusion is performed on the output images of streams with different resolutions, which can form a high-resolution network structure. Finally, the highest resolution of the feature map is used to predict facial landmarks. It maintains high-resolution representations through the whole process for spatially precise heatmap estimation. The network mainly includes parallel multi-resolution convolution, repeated multi-resolution fusion and regression heatmap.

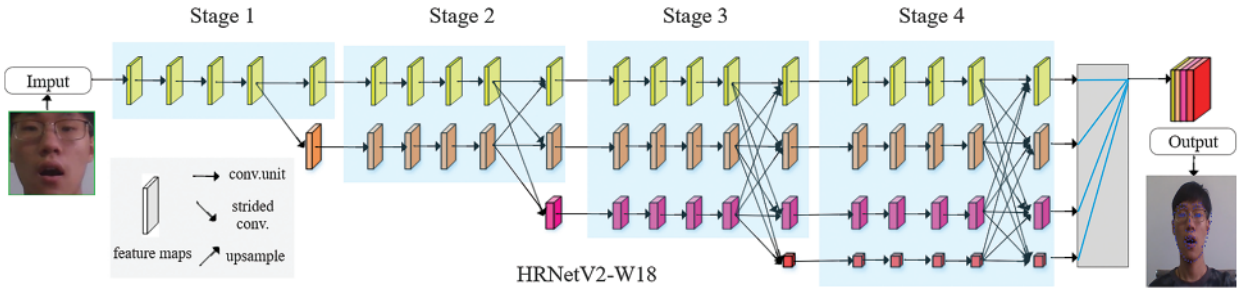

Figure 6: HRNetV2-W18 structure

4.2.1 Parallel Multi-Resolution Convolution

Starting from the input high-resolution image information as the first stage, the streams from high-resolution images to low-resolution images are added layer by layer to form a parallel integration and connection stage, when the network depth gradually increases. The main body of the network consists of four parallel streams, composed of two stride-2 3×3 convolutions, whose resolution is reduced by half in turn and the corresponding number of channels is doubled. Therefore, the resolutions for the parallel streams of a later stage consist of the resolutions from the previous stage, and an extra lower one. An example network structure, containing 4 parallel streams, is logically as follows:

$$\begin{array}{ccccccc} N_{11} & \rightarrow & N_{21} & \rightarrow & N_{31} & \rightarrow & N_{41} \\ & \searrow & & & & & \\ & & N_{22} & \rightarrow & N_{32} & \rightarrow & N_{42} \\ & & & \searrow & & & \\ & & & & N_{33} & \rightarrow & N_{43} \\ & & & & & \searrow & \\ & & & & & & N_{44} \end{array} \tag{6}$$

where N_{sr} is a sub-stream in the sth stage, and r is the resolution index. The resolution index of the first stream is $r = 1$. The resolution of index r is $\frac{1}{2^{r-1}}$ of the resolution of the first stream.

4.2.2 Repeated Multi-Resolution Fusion

The process of multi-resolution fusion in the parallel network is shown in Fig. 7. For feature maps larger than the current scale, a convolution layer with a stride-2 3×3 convolution is used for down-sampling processing. After passing through the convolution layer, the side length of feature maps becomes half of the original and the area becomes one quarter of the original. For feature maps smaller than the current scale, up-sampling operation is required, the side length is doubled and the area is quadrupled. Firstly, interpolation method is used to expand the resolution. Then, the convolution layer with 1×1 convolution is used to change the number of channels. Finally, the features from different sources are summed to obtain the fused features.

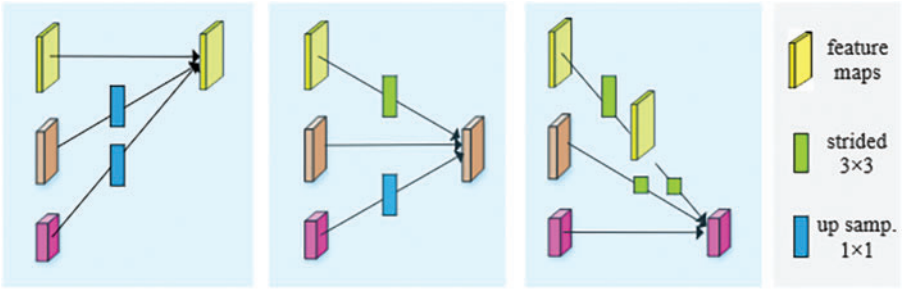

Figure 7: Fusion process of high, medium and low resolution respectively

The exchange units introduced across parallel streams in multi-resolution fusion is used to make each stream repeatedly receive information from other parallel streams. Taking the third stage as an example, it is divided into several exchange units composed of three parallel convolutional units and

one exchange unit across the parallel units. The scheme of exchanging information is as follows:

$$\begin{array}{ccccccccccc} C_{31}^{1} & \searrow & & \nearrow & C_{31}^{2} & \searrow & & \nearrow & C_{31}^{3} & \searrow & \\ C_{32}^{1} & \rightarrow & \varepsilon_{3}^{1} & \rightarrow & C_{32}^{2} & \rightarrow & \varepsilon_{3}^{2} & \rightarrow & C_{32}^{3} & \rightarrow & \varepsilon_{3}^{3} \\ C_{33}^{1} & \nearrow & & \searrow & C_{33}^{2} & \nearrow & & \searrow & C_{33}^{3} & \nearrow & \end{array} \tag{7}$$

where C_{sr}^{b} represents the convolution unit in the rth resolution of the bth block in the sth stage, and ε_{s}^{b} is the corresponding exchange unit. Formula (7) corresponds to Fig. 7.

Each output is a collection of input mappings, as shown in the following formula:

$$\mathrm{Y}_{k} = \sum_{i=1}^{8} a(X_i, k) \tag{8}$$

The exchange unit across stages has an extra output map, as shown in the following formula:

$$\mathrm{Y}_{s+1} = a(Y_s, s+1) \tag{9}$$

where s is the number of parallel streams, the inputs are s response maps: $\{X_1, X_2, \ldots, X_S\}$, the outputs are s response maps: $\{Y_1, Y_2, \ldots, Y_s\}$, whose resolutions and widths are the same to the input. The function $a(X_i, k)$ consists of up-sampling or down-sampling X_i from resolution i to resolution k.

4.2.3 Heatmap Estimation

Heatmap regression firstly extracts features of face images, and then learns the features information from HRNetV2-W18 model. Finally convolution layer is added to the last layer of the model, and the positions of facial landmarks estimated by regressors are converted into high resolution heatmaps. The improvement of HRNetV2 model compared with HRNetV1 mainly lies in the fact that the output of the final feature map integrates the information from low-resolution feature maps, which can improve the accuracy of facial landmarks recognition task, as shown in Fig. 8 below.

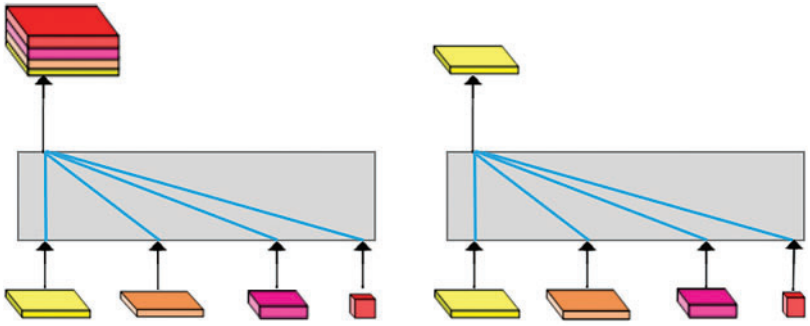

Figure 8: Different types of HRNet output layers

The network is applied to a task of facial landmarks recognition. For face image P, the network can obtain N heatmap $H(P)$, where N is the total number of facial landmarks. The heatmap based on landmarks recognition of the network adopts the gaussian distribution principle that is the maximum value in the position of the heatmap, decodes the predicted position of each landmark from the corresponding heatmap, and output the coordinate of facial landmarks. As shown in the following

formula:

$$\mathrm{L(i)} = argmax H^{i}(P) \tag{10}$$

where i is the heatmap index corresponding to facial landmark, $L(i)$ gives the coordinates of the ith landmark.

5 Experimental Results and Application

In this section, extensive experiments and analyses are carried out to prove the robustness and effectiveness of the proposed method. Firstly, the following paragraphs describe the datasets, training details and evaluation metric. Secondly, in terms of performance and metrics, the MTHR-Face is compared with other algorithms and its results are analyzed. Finally, the method in this paper is used to capture facial expression and generate animation in real-time.

5.1 Datasets

WIDER FACE [25] is a benchmark database for face detection including 32,203 images and 393,703 labeled face bounding boxes. 70% of the images are taken as a training set and the remaining as a test set.

The Wider Facial Landmarks in the Wild (WFLW) [19] is considered the most challenging database, which contains 10,000 faces (7500 for training and 2500 for testing) with 98 fully manual annotated facial landmarks. This database also features rich attribute annotations in terms of pose, expression, illumination, makeup, occlusion and blur , which can effectively evaluate the robustness of the algorithm for large angle posture and complex expression.

5.2 Training Details

The experiment uses Python3.7 and is compiled in PyCharm integrated development environment. The two models of MTHR-Face algorithm, improved MTCNN and HRNetV2-W18, were trained independently. Both networks were implemented in PyTorch. In this paper, the MTHR-Face algorithm is implemented on Windows10 operating system with NVIDIA GTX2080Ti (16 GB) GPU and an Intel(R) Core(TM) i7-8700 CPU @ 3.20 GHz.

For improved MTCNN, P-Net, R-Net and O-Net are trained separately. Minibatch size is set to 256 and learning rate is 1e-3. In the process of network training, loss functions determined by Eqs. (1)–(3) are adopted for the three main tasks of MTCNN to carry out network training. And Intersection-over-Union (IoU) ratio is calculated by using real face coordinate area. If IoU value is greater than 0.65, it is a positive sample. If IOU value is less than 0.3, it is a negative sample. IoU value between 0.4 and 0.65 are considered to contain only partial face information. Negative and positive samples are used for face classification tasks, positive and partial faces are used for bounding box regression.

The HRNetV2-W18 model is trained that the input images are cropped by face candidate windows of improved MTCNN and resized to 256×256 resolution. Data augmentation is applied by random flipping, rotation(between $\pm 30^{\circ}$), increasing gaussian noise and color enhancement in WFLW database to improve the generalization ability of the network and the robustness of the model, as shown in Fig. 9 below. The model is optimized using the Adam optimizer with an initial learning rate of 0.0001 and the batch size was set to 16. In total, training is applied of 300 epochs.

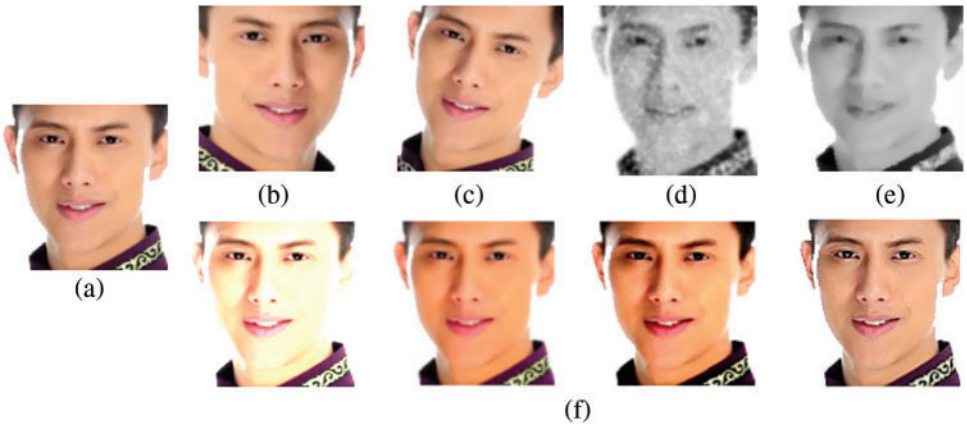

Figure 9: Example diagram of image enhancement method for WFLW database. ((a) A sample cropped face image from the WFLW database; (b) Image flipping; (c) Image rotation; (d) Image Gaussian noise; (e) Image median filtering; (f) One group is image color enhancement)

5.3 Evaluation Metric

In this paper, we use the normalization mean error (NME), the failure rate (FR), the area under the curve (AUC) and the cumulative error distribution (CED) curve of samples to measure the facial landmark location error.

NME is a widely used metric to evaluate the performance of facial landmark recognition. Error of each landmark is calculated this way and then averaged to get the final result. The formula is as follows:

$$\text{NME} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left\|\hat{x}_s - x_s\right\|_2}{dist\_between\_eyes} \tag{11}$$

where N indicates the number of facial landmarks, $\hat{x}_s$ indicates the predicted value of the facial landmarks, x_s indicates the actual value of the facial landmarks. For the WFLW database, *dist_between_eyes* indicates the distance between the outer eye corners ("inter-ocular").

FR is calculated based on the NME value. For the WFLW database, the threshold ε is set to 0.1. When the NME of an image is larger than 0.1, this case is deemed a failure of facial landmark recognition. We derive the FR from the rate of failures in a test set.

AUC provides another insight into a design of facial landmarks recognition. Basically, it can be deduced from CED curve that a non-negative curve is formed by plotting the curve from zero to the threshold for FR, under which the area is calculated to be AUC. The formula is as follows:

$$\text{CED} = \frac{N_{e\leq l}}{N} \tag{12}$$

where $N_{e\leq l}$ represents the number of images whose error l is not less than e. Performance is higher when the CED curve is correspondingly higher. And CED curve evaluation indexes are widely used in benchmark database for facial landmark recognition.

5.4 Experimental Results and Analysis

Since the WFLW database meets the diversity of expressions and contains various type of challenge, the proposed method in this paper is compared with LAB algorithm and HRNetV2-W18 model alone from the final output subjective result images and the above evaluation metrics, which can comprehensively evaluate the robustness of MTHR-Face.

The LAB is much more computational expensive due to a network architecture using eight-stacked hourglass modules than MTHR-Face and HRNetV2-W18 model. And the hourglass network is adopted to first reduce the resolution and then increase the resolution, whereas HRNetV2-W18 is realized by gradually adding high-to-low resolution streams to connect the multi-resolution streams in parallel, and repeated multi-scale fusions are performed, which can maintain the high-resolution information of face feature maps and is accurate in space. LAB that a low-to-high resolution network structure has the risk of losing characteristic information, as shown in Fig. 10b. When the human face makes complex expression changes, HRNetV2-W18 alone is used in Fig. 10c where some landmarks are recognized beyond the target face region. However, the MTHR-Face method proposed in this paper combines HRNetV2-W18 with the improved MTCNN, which not only effectively maintains the spatial information of the face, but also can recognize facial landmarks more accurately under different levels of expressions to achieve real virtual animation effects for the subsequent. Fig. 10 below shows the comparison results of test faces in facial landmarks recognition under different algorithms. It can be seen that the MTHR-Face method in Fig. 10d can recognize facial landmarks more accurately under different expressions, such as normal and amazed, because the improved MTCNN can adjust the face posture, detect the face, cut the face image based on the face candidate windows and then input into HRNetV2-W18 model, which can further reduce the difficulty of landmarks regression task and make the recognized facial landmarks more convergence.

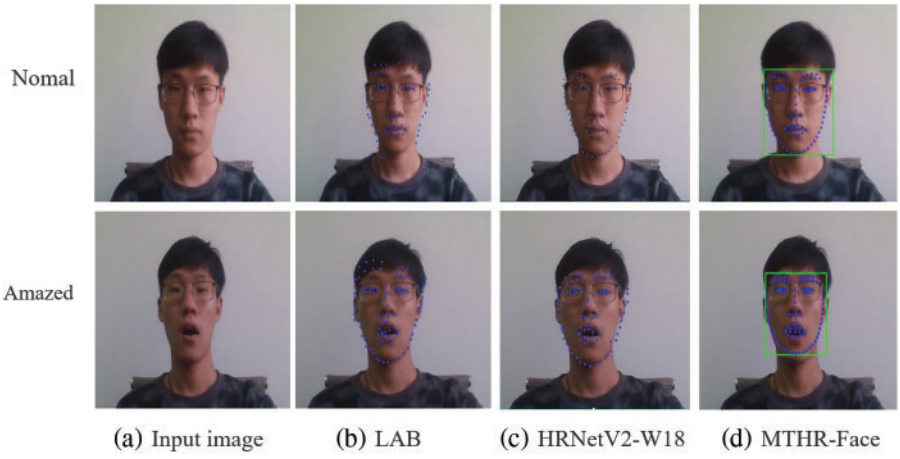

Figure 10: Comparison diagram of the results of each algorithm under different levels of expressions

To further evaluate the performance of the MTHR-Face method, Tab. 1 below illustrates NME (lower is better), FR (lower is better), AUC (higher is better) and CED curve on the testset and six subsets of WFLW among the MTHR-Face, the HRNetV2-W18 alone and the LAB algorithm. The results show that MTHR-Face performs better than LAB and HRNetV2-W18 alone by a significant margin on every subset. LAB is weak in recognizing facial landmarks of extreme diversity samples on WFLW, such as big pose, exaggerated expressions and heavy occlusion. However, when combined with the improved MTCNN, HRNetV2-W18 decreases NME from 4.60% to 4.39%. Note that MTHR-Face

still outperforms HRNetV2-W18 alone in all other metrics, which is much more beneficial to facial landmarks recognition. Compared with other method, the innovations proposed in this paper exhibit a certain improvement for each subset of the WFLW database. These results demonstrate that the method improve the problem of low accuracy and deviation caused by diverse expressions or complex background. Besides, the CED curve in Fig. 11 shows that MTHR-Face curve is higher than the rest two between 0.02 and 0.1, which means the performance of the proposed method is significantly better than that of other algorithms in WFLW testset and it has certain robustness to extreme changes under different degrees of expressions.

Table 1: Comparison in terms of NME (lower is better), FR (lower is better) and AUC (higher is better) on WFLW testset and its six subsets (98 landmarks)

Metric	Method	Fullset	Pose	Expression	Illumination	Makeup	Occlusion	Blur
NMS (%)	LAB [19]	5.27	10.24	5.51	5.23	5.15	6.79	6.32
	HRNetV2–18	4.60	7.86	4.79	4.57	4.26	5.42	5.36
	MTHR-Face	**4.39**	**7.38**	**4.61**	**4.32**	**3.76**	**5.11**	**4.90**
FR (%)	LAB [19]	7.56	28.83	6.37	6.73	7.77	13.72	10.74
	HRNetV2–18	3.24	16.8	2.23	3.01	1.94	5.98	4.53
	MTHR-Face	**3.02**	**14.8**	**2.19**	**2.83**	**1.86**	**5.64**	**4.12**
AUC@0.1	LAB [19]	0.5323	0.2345	0.4951	0.5433	0.5394	0.4490	0.4630
	HRNetV2–18	0.5528	0.2860	0.5360	0.5616	0.5763	0.4862	0.4886
	MTHR-Face	**0.5871**	**0.3242**	**0.5579**	**0.5692**	**0.5957**	**0.5388**	**0.5536**

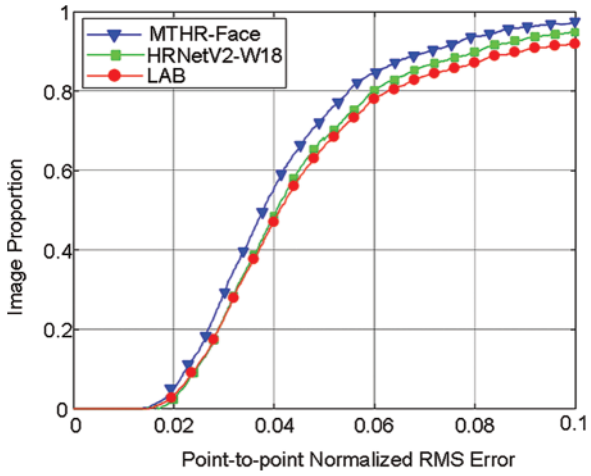

Figure 11: CED curves of different algorithms using WFLW database

5.5 Animation Generation

After collecting data from 98 facial landmarks based on the MTHR-Face in this paper, Support Vector Machine (SVM) [26] is used to train the strength estimation model of facial action unit (AU) to extract the parameters [27], which can obtain the classification and regress effect of facial expression. Finally, the expression parameters are acquired by establishing the mapping relationship between AU and Blendshape expression bases to realize facial expression capture and animation generation.

Blendshape [28] is a linear combination of natural and other facial shapes and controls the facial expression of 3D models through semantic weight. Based on the division method of Facial Action Coding System, the specific facial expression is represented by E:

$$\mathrm{E} = E_0 + \sum_{i=0}^{17} e_i(E_i - E_0) \tag{13}$$

Where E_0 is expression base, e_i is the weighted coefficient of corresponding expression and the facial expression module contains 18 data of facial action unit.

In the process of expression animation drive, we firstly use Maya [29] to model the test virtual character in Fig. 12 and bind the facial skeleton, then design the BlendShape expression controller of the character model and create the corresponding AU shape expression base on the virtual character's face. Secondly, AU parameters extracted from real faces after normalization are mapped to Morph Targets controller in Unity3D engine. Finally, the mapping relationship is used to drive the facial expression changes of virtual characters to realize the man-computer facial expression interaction [30].

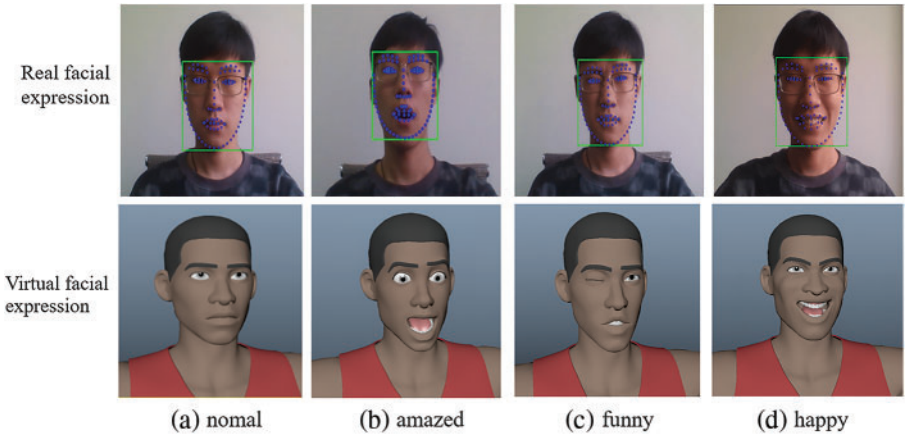

Figure 12: Real-time expression generation animation effects

The MTHR-Face model is used to capture the changes of facial features in real-time, and the virtual character shaped by Maya is selected as the test target model. Then facial expression captured in real-time is mapped to generate facial synchronous expression animation. Different emotion-driven animation effects are obtained as shown in Fig. 12. Ensuring real-time performance, the average frame rate of generated animation reaches 20fps. It can be seen that the MTHR-Face can accurately capture the user's facial expression and drive the virtual character expression animation in real-time.

6 Conclusion

MTHR-Face model based on two-stage neural network is proposed for facial expression capture. The method leverages the best advantages of the improved MTCNN and HRNetV2-W18 model. Benefiting from robust face detection of the improved MTCNN, the cropped images are input into HRNetV2-W18 model to obtain 98 high-resolution facial landmarks, which not only effectively maintains the spatial information of the face, but also can recognize facial landmarks more accurately under different levels of expressions and complex background. Finally, based on the movement of the recognized landmarks, the expression parameters are then transmitted to the Unity3D engine to drive the virtual character's face for generation of real-time expression animation. Experimental results

show that, the method achieves more outstanding performance than others in all metrics on WFLW database and can better improve the accuracy of facial landmark recognition. Especially compared with HRNetV2-W18 alone, the method's NME is reduced from 4.60% to 4.39%. MTHR-Face can accurately capture facial expression and generate animation. Future work will consist in training other databases to verify the MTHR-Face, or possibly other representations and tasks, such as head pose estimation and face texture feature extraction, to get rich emotional communication and interactive between user and avatar in shared virtual space.

Acknowledgement: We would like to thank the anonymous reviewers for their valuable and helpful comments, which substantially improved this paper. At last, we also would also like to thank all of the editors for their professional advice and help.

Funding Statement: This research was funded by College Student Innovation and Entrepreneurship Training Program, grant number 2021055Z and S202110082031, the Special Project for Cultivating Scientific and Technological Innovation Ability of College and Middle School Students in Hebei Province, Grant Number 2021H011404.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] X. J. Wu and G. L. Ju, "A markerless facial expression capture and reproduce algorithm," *Acta Electronica Sinica*, vol. 44, no. 9, pp. 2141–2147, 2016.

[2] H. Jiang, R. Jiao, D. Wu and W. Wu, "Emotion analysis: Bimodal fusion of facial expressions and eeg," *Computers, Materials & Continua*, vol. 68, no. 2, pp. 2315–2327, 2021.

[3] Z. Xia, L. Jiang, X. Ma, W. Yang, P. Ji *et al.,* "A Privacy-preserving outsourcing scheme for image local binary pattern in secure industrial internet of things," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 629–638, 2019.

[4] A. Siyaev and G. S. Jo, "Towards aircraft maintenance metaverse using speech interactions with virtual objects in mixed reality," *Sensors*, vol. 21, no. 6, pp. 2066–2086, 2021.

[5] X. H. Huang, W. H. Deng, H. F. Shen, X. B. Zhang and J. P. Ye, "PropagationNet: Propagate points to curve to learn structure information," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 7265–7274, 2020.

[6] D. Datta, P. K. Maurya, K. Srinivasan, C. Chang, R. Agarwal *et al.,* "Eye gaze detection based on computational visual perception and facial landmarks," *Computers, Materials & Continua*, vol. 68, no. 2, pp. 2545–2561, 2021.

[7] X. R. Zhang, T. Xu, W. Sun and A. G. Song, "Multiple source domain adaptation in micro-expression recognition," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 8371–8386, 2021.

[8] D. Sibbing, M. Habbecke and L. Kobbelt, "Markerless reconstruction and synthesis of dynamic facial expressions," *Computer Vision and Image Understanding*, vol. 115, no. 5, pp. 668–680, 2011.

[9] C. Cao, Y. Weng, S. Zhou, Y. Tong and K. Zhou, "FaceWarehouse: A 3D facial expression database for visual computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 2, pp. 413–425, 2013.

[10] C. Cao and Y. Weng, "3D shape regression for real-time facial animation," *ACM Transactions on Graphics*, vol. 32, no. 41, pp. 1–10, 2013.

[11] T. Weise, S. Bouaziz and H. Li, "Realtime performance-based facial animation," *ACM Transactions on Graphics*, vol. 30, no. 77, pp. 1–10, 2011.

[12] S. Laine, T. Karras and T. Aila, "Production-level facial performance capture using deep convolutional neural networks," in *SCA '17: The ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, Los Angeles, United States, pp. 1–10, 2017.
[13] J. Cheng, R. M. Xu, X. Y. Tang, V. S. Sheng and C. T. Cai, "Locating facial features with an extended active shape model," in *European Conf. on Computer Vision*, Marseille, France, pp. 504–513, 2008.
[14] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.
[15] Z. H. Feng, J. Kittler and W. Christmas, "Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting," in *2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 2481–2490, 2017.
[16] Y. Sun, X. Wang and X. Tang, "Deep convolutional network cascade for facial point detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Portland, OR, USA, pp. 3476–3483, 2013.
[17] E. Zhou, H. Fan and Z. Cao, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade," in *2013 IEEE Int. Conf. on Computer Vision Workshops*, Sydney, NSW, Australia, pp. 386–391, 2013.
[18] A. Newell, K. Yang and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*, Amsterdam, The Netherlands, pp. 483–499, 2016.
[19] W. Y. Wu, C. Qian and S. Yang, "Look at boundary: A boundary-aware face alignment algorithm," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 2129–2138, 2018.
[20] J. Yang, Q. S. Liu and K. H. Zhang, "Stacked hourglass network for robust facial landmark localisation," in *2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Honolulu, HI, USA, pp. 2025–2033, 2017.
[21] K. P. Zhang and Z. P. Zhang, "Joint face detection and alignment using multi-task cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.
[22] Y. Chen, H. Fan, B. Xu, Z. Yan, Y. Kalantidis *et al.,* "Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution," in *2019 IEEE/CVF International Conference on Computer Vision*, Seoul, Korea (South), pp. 63–72, 2019.
[23] J. Wang, K. Sun, T. Cheng, B. Jiang and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2019.
[24] K. Sun, B. Xiao, D. Liu and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 5686–5696, 2019.
[25] S. Yang, P. Luo, C. C. Loy and X. Tang, "WIDER FACE: A face detection benchmark," in *2016 IEEE Conf.on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 5525–5533, 2016.
[26] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2930–2940, 2013.
[27] Y. Q. Li, J. X. Chen, Y. P. Zhao and J. Qiang, "Data-free prior model for facial action unit recognition," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 127–141, 2013.
[28] K. Prabhu, S. SathishKumar, M. Sivachitra, S. Dineshkumar and P. Sathiyabama, "Facial expression recognition using enhanced convolution neural network with attention mechanism," *Computer Systems Science and Engineering*, vol. 41, no. 1, pp. 415–426, 2022.
[29] J. Liu, Y. Li and J. P. Zhu, "3D virtual human animation generation based on dual-camera capture of facial expression and human pose," *Journal of Computer Applications*, vol. 41, no. 3, pp. 839–844, 2021.
[30] D. J. Huang, Y. Q. Yao, W. Tang and Y. D. Ding, "Facial tracking and animation for digital social system," in *VRCAI '2018: Int. Conf. on Virtual Reality Continuum and its Applications in Industry*, Tokyo, Japan, pp. 1–8, 2018.