Hindawi
IET Biometrics
Volume 2023, Article ID 6636386, 14 pages
https://doi.org/10.1049/2023/6636386

Research Article

Adaptive Weighted Face Alignment by Multi-Scale Feature and Offset Prediction

Jingwen Li, Jiuzhen Liang, Hao Liu, and Zhenjie Hou

School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou 213164, Jiangsu, China

Correspondence should be addressed to Jiuzhen Liang; jzliang@cczu.edu.cn

Received 25 June 2023; Revised 15 September 2023; Accepted 7 November 2023; Published 6 December 2023

Academic Editor: Jun Wan

Copyright © 2023 Jingwen Li et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Traditional heatmap regression methods have some problems such as the lower limit of theoretical error and the lack of global constraints, which may lead to the collapse of the results in practical application. In this paper, we develop a facial landmark detection model aided by offset prediction to constrain the global shape. First, the hybrid detection model is used to roughly locate the initial coordinates predicted by the backbone network. At the same time, the head rotation attitude prediction module is added to the backbone network, and the Euler angle is used as the adaptive weight to modify the loss function so that the model has better robustness to the large pose image. Then, we introduce an offset prediction network. It uses the heatmap corresponding to the initial coordinates as an attention mask to fuze with the features, so the network can focus on the area around landmarks. This model shares the global features and regresses the offset relative to the real coordinates based on the initial coordinates to further enhance the continuity. In addition, we also add a multi-scale feature pre-extraction module to preprocess features so that we can increase feature scales and receptive fields. Experiments on several challenging public datasets show that our method gets better performance than the existing detection methods, confirming the effectiveness of our method.

1. Introduction

Facial landmark detection, also known as face alignment, is an important part of face correlation research in computer vision. In contrast to the landmarks on a human body, a rigid body such as a face has constant relative positions of landmarks that must be calibrated, such as eyebrows, eyes, noses, lips, etc., based on prior knowledge. Many works in face correlation research rely on the facial landmark detection technology, such as face recognition [1], face animation synthesis [2], frontal face reconstruction [3], etc.

In recent years, the mainstream methods for the facial landmark detection have been divided into two categories: coordinate regression-based and heatmap regression-based. Both methods have achieved good results in the detection effect. However, due to the different prediction principles, the advantages and disadvantages of these two regression methods are also obvious. For the fully connected direct coordinate regression method, image features are extracted through various convolution structures first, the extracted features are integrated, and the coordinates of landmarks are directly regressed through the fully connected layer. The advantages of this method are fast training speed and end-to-end regression. However, the fully connected regression method is extremely dependent on the spatial distribution of the input image, so it is very short of spatial generalization ability, and the accuracy is not high compared with the heatmap regression methods.

The method of heatmap regression, and its prediction principle is completely different from the coordinate regression methods. First, image features are extracted by up–down sampling to generate a probabilistic heatmap. The pixel index corresponding to the probability peak is obtained by the Argmax method, which is the position coordinate of the predicted landmarks. The advantages of this method are relatively higher accuracy and accurate local position prediction. A defect of this model is slow training speed and it is not end-to-end. To take into account factors such as memory consumption, the number, and the training speed, heatmap regression-based methods usually take a quarter of the size of the input size as the dimensions of the output probability heatmaps. The coordinates are integers, and the precision of

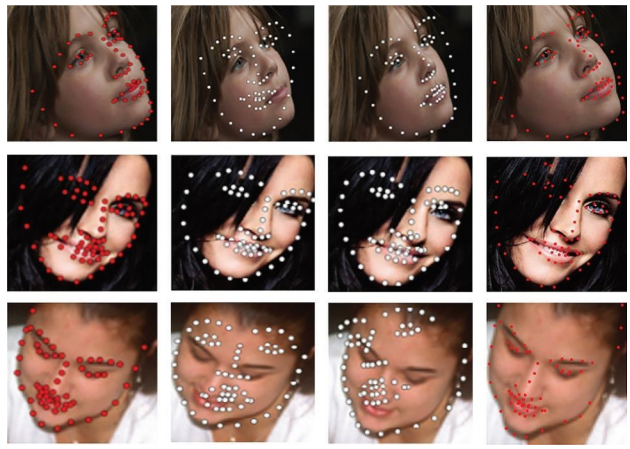

FIGURE 1: The test results of our method which compared with those of other methods in an unrestricted field environment. From left to right, each column is the (a) ground truth, (b) CFSS [4], (c) LBF [5], and (d) the results of our method.

the ground truth does not match, resulting in the lower bound theory of error and inevitable.

In general, the current commonly used methods of face alignment have poor performance in complex environments such as large poses, exaggerated expressions, occlusion, different lighting, makeup, etc. Figure 1 shows the comparison of the effects of various detection methods under several different interference conditions. As shown in Figure 1, each row from top to bottom is the model detection results under large poses, occlusion, and blur. From left to right, each column is the ground truth, CFSS [4], LBF [5], and the results of our method. Based on the principles and characteristics of the two commonly used methods mentioned above, we summarize the following defects: (1) heatmap regression pays more attention to the local features and lacks spatial relationship constraints between adjacent landmarks, which will destroy the continuity and global structure of face shape when doing feature matching. (2) There is an uneven distribution of samples in the dataset. A large number of frontal images will affect the network's feature learning of large pose images, leading to the problem of data imbalance. (3) The feature scale and receptive field range obtained during the feature extraction by simple convolution operation are limited, which may lead to a decrease in detection accuracy.

Considering the above problems and defects, we propose an adaptive weighted face alignment model based on multi-scale feature and offset prediction (AWMOP). Considering the accuracy of the traditional coordinate regression method is poor and the theoretical error lower bound problem will inevitably occur in heatmap regression, this paper adopts the hybrid detection model [6] combining heatmap regression and coordinate regression as the backbone network. Then, add an offset prediction network (OPN) by regression to achieve the global shape constraint. Specifically, at the early stage of the model, a multi-scale feature pre-extraction (MFP) module is embedded to increase the receptive field and integrate more features of the different levels. Meanwhile, we also add the three-dimensional Euler angle prediction module (HPP) to the initial coordinate prediction to obtain face rotation information, which is used to adjust the loss function. An image with a large change of pose is punished with a heavier weight, while an image with a smaller change is penalized. After obtaining the initial coordinates of the landmarks of the face through the first stage, the corresponding two-dimensional Gaussian heatmap is regenerated as a mask. Then, put the heatmap mask into the attention module of the offset prediction network in the second stage. The coordinate offsets are generated by regression, and the spatial constraints between the coordinates of adjacent blocks are optimized to achieve the purpose of global constraints. As shown in Figure 1, our method still performs well in unconstrained field environments.

In general, the specific contributions of our work are summarized as follows:

(1) We introduce an offset prediction network to optimize the global spatial constraints, refine the location of landmarks, and add an attention module to focus on the features around landmarks, which improves our understanding of the spatial context relationship between landmarks and allows us to further optimize their coordinates.

(2) In the hybrid detection model combining heatmap regression and coordinate regression, we embedded the head rotation attitude prediction module as the adaptive weight of the loss function to solve the problem of data imbalance.

(3) We develop a feature pre-extraction module to increase the receptive field, increase the number of feature maps to obtain more features, and integrate features of different levels to obtain more accurate effects.

(4) Our method has achieved good results on popular universal facial landmark detection datasets, including 300 W [7], COFW [8], and WFLW [9], demonstrating its effectiveness and robustness.

2. Related Work

In recent years, computer vision has developed rapidly and has been widely used, such as image processing, image segmentation, and other general directions. In the field of image processing, facial image research has also experienced a series of development and improvements. Facial landmark detection is a very important basic work in the field of facial image research. Due to its fundamental importance, it has been intensively studied in the recent years. There are currently two main categories of mainstream methods: coordinate-based regression methods and heatmap-based regression methods.

Coordinate-based regression method: this method mainly learns the facial features of the face directly, and regresses the coordinate information of the landmarks directly through the full connection mode. For example, a coarse to fine facial landmark detection algorithm proposed by Zhou et al. [10], which divided facial landmarks into internal landmarks and boundary landmarks. The final coordinates of the face were obtained by overlaying the outputs of two cascaded CNN for two parts landmarks. Zhang et al. [11] first used multitask learning combined with CNN to detect facial landmarks. Multiple auxiliary tasks such as smile detection, posture detection to assist the main task, making the convergence faster and higher accurate. Wu et al. [12] used Vanilla CNN to cluster K categories based on the features obtained from the fully connected layer, and divided the training images into different categories. The author used images with similar features to train the corresponding regressors, and ultimately achieved good results. Li et al. [13] proposed a new topological adaptive deep map learning method to obtain accurate facial landmarks. Guo et al. [14] used a lightweight network model to regress the parameters of 3DMM and dynamically combine WPDC and VDC loss functions, thus accelerating the speed of fitting. Zhang et al. [15] proposed a two-stage cascade regression alignment model, which generated rough initial shapes from aligned salient pole shapes.

Heatmap-based regression method: this method mainly obtains the likelihood heatmap for each key point and then obtains the coordinates of the landmarks in the heatmap by Argmax. For example, Valle et al. [16] first calculated the probability map of each feature point through the CNN model to get the feature point position, and then calculated the pose matrix to further improve the detection accuracy. Ullah et al. [17] proposed a facial alignment algorithm called Double Attention Spatial Perception Capsule Network (DSCN). The authors utilized the hourglass capsule network and adaptive local constraint dynamic routing algorithm to capture the spatial positional relationships of features, and obtained a facial boundary heatmap to improve accuracy. Wan et al. [18] proposed an implicit multi order correlation model and an explicit probability based boundary adaptive regression (EPBR) method to enhance global shape constraints. Huang et al. [19] learned the problem of error deviation in the face of alignment. This method used an hourglass network as the backbone, combined with anisotropic directional loss and anisotropic attention module. Bulat and Tzimiropoulos [20] introduced a method divided into two parts. First, providing a confidence score for the position of each facial marker by convolution. Then, combining the confidence heatmap and high-resolution features to regress the coordinates. Yang et al. [21] enhanced the representation ability of local features and global context features by mixing the dual attention mechanism. Xie et al. [22] mapped the boundary heatmap to the landmark heatmap to improve the conversion efficiency, instead of directly using the boundary heatmap to return the coordinates of facial landmarks.

At the same time, in addition to these two conventional detection methods, there are also many other detection models based on the above two methods. For example, Park and Kim [23] combined a coordinate regression network and heatmap regression network with spatial attention to complement the defects of heatmap regression and coordinate regression to deal with the occlusion problem. Meanwhile, in recent years, transformer has also shown good results in facial alignment work. Xia et al. [24] used transformer to learn the internal relationships between coordinate points, and used a coarse-to-fine iterative framework to optimize coordinate positions until convergence. Li et al. [25] learned the structured relationships between landmarks through the self-attention of the transformer. It adopted a cascade refinement process for coordinate optimization, extracting relevant image features around the target point for coordinate prediction. Additionally, it refined landmark positions and image features using a new decoder.

Considering the advantages and disadvantages of the above detection methods, we propose a hybrid landmark detection model assisted by offset prediction. The method combines heatmap regression and coordinate regression and incorporates a feature pre-extraction module and 3D head pose prediction module to improve the accuracy of initial coordinates. At the same time, an offset prediction module is introduced, which combines the attention mechanism to predict the offset and further refines the coordinates of landmarks.

3. Main Work

In this section, we detail our AWMOP, whose flowchart is shown in Figure 2. The network consists of two stages. The first stage is a hybrid detection model for predicting initial coordinates, and the second stage is an offset prediction network for predicting coordinate offsets. First, the input images get the face features through the MFP module; the feature map enters the initial coordinate prediction module to get the initial coordinates, and then gets the final results through the offset prediction network. The MFP module in the first stage is shown in Section 3.1, the initial prediction network of our model in the first stage is detailed in Section 3.2, and the structure of the offset prediction network in the second stage is finally introduced in Section 3.3.

3.1. Multi-Scale Feature Pre-Extraction Model. Due to the characteristics of the face, the extraction of local features of the landmark location region of the face is particularly important, especially when dealing with the face samples with large expression changes or makeup interference. In

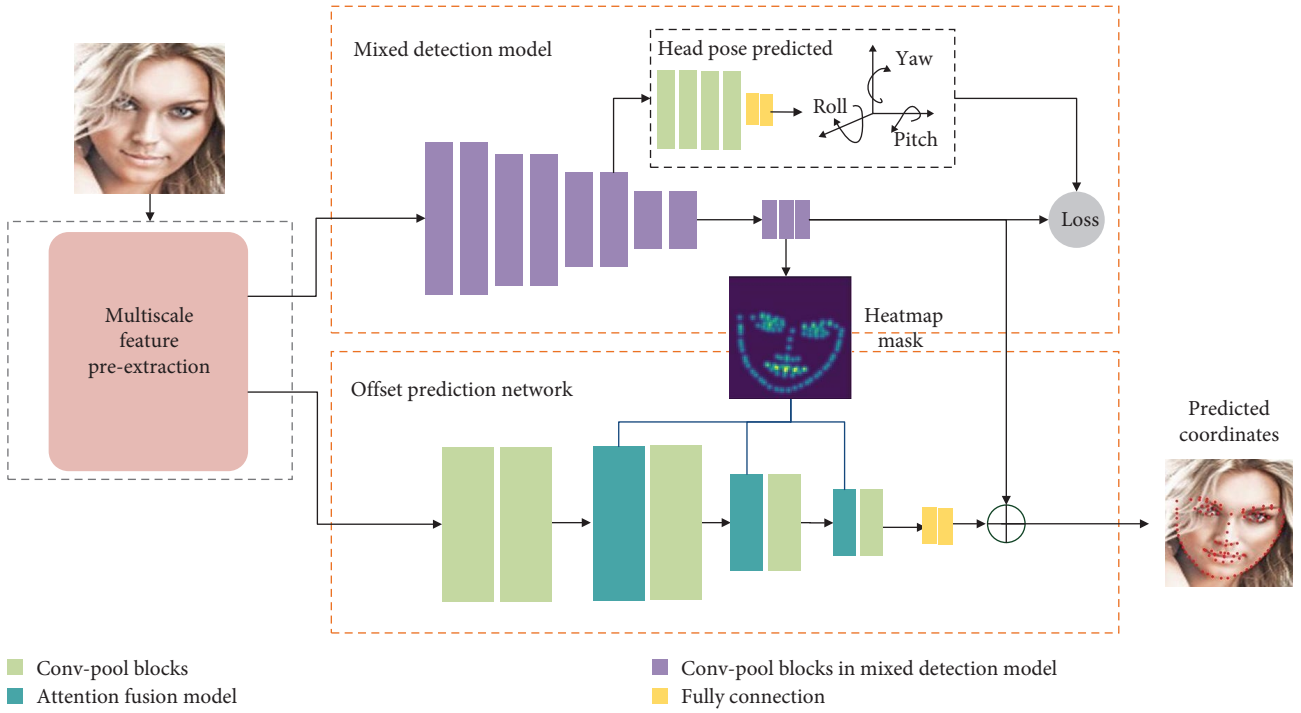

FIGURE 2: The overall architecture of the network. The network consists of two stages. The first stage is a hybrid detection model for predicting initial coordinates, and the second stage is an offset prediction network for predicting coordinate offsets.

recent years, several works have shown that multi-scale feature extraction operations are necessary before the network can convolve the original image to extract features. The bidirectional residual correction network for processing multi-scale feature information extracted from different feature layers proposed by Tang et al. [26] can effectively learn shallow and deep features, supplement semantic and detail information, and better complete fuzzy image detection tasks. DeFusionNet [27] propagates the fuzed shallow features to the deep layers to refine the details of the detected defocused blurry areas, and propagates the fuzed semantic features to the shallow layers to help better locate the blurry areas. It repeatedly fuzes and refines multi-scale deep features to improve the effectiveness of defocused blurry detection. Therefore, we design a MFP module to complete the feature preprocessing before network detection.

In order to obtain larger receptive fields and more local features, we want more multi-scale facial features. Therefore, we introduce the feature pre-extraction module which is divided into four branches. As shown in Figure 3, $H \times W \times C_I/C_O$ represents the height, width, and corresponding channel number of the input and output feature maps, respectively, and the channel number of each branch is 1/4 of the output channel number. These four branches run in parallel, and then feature join. In the first branch, we use convolution with a convolution kernel of 1×1 for feature extraction. In the second branch, convolution operations with kernels of 1×1 and 3×3 are adopted, respectively, where the number of channels corresponding to the first convolution is 1/8 of the total number of output channels. In the third branch, first, the maximum pooling layer with a

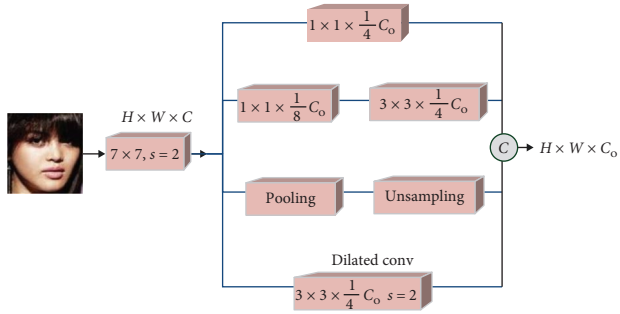

FIGURE 3: Multi-scale feature pre-extraction module architecture.

pooling window of 2×2 and a step size of 2 is used to realize downsampling, and then the upsampling operation is carried out to adjust the feature map to the original size. The advantages of this method are that the number of feature maps can be increased, and the sample features of more levels can be extracted and fuzed. So, we can get accurate effects in local details. In the last branch, we employ cavity convolution with a kernel of 3×3 to further expand the receptive field. By connecting the output feature maps of the four branches, we can obtain the multi-scale features of the input image. Feature pre-extraction before the whole network can effectively extract local face information, and it has a better learning effect for sample features with exaggerated expressions and makeup interference.

3.2. Initial Prediction Networks. The traditional coordinate regression has many unaligned points, making it insensitive to facial expression changes, such as blinking and crooked mouths. Heatmap regression, however, has a poor effect on

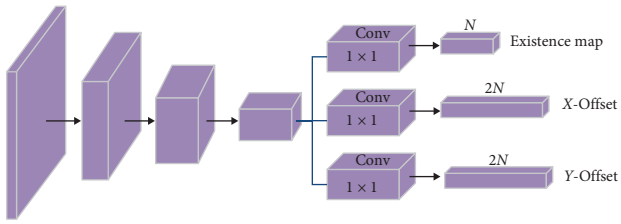

Figure 4: Structure of hybrid detection model.

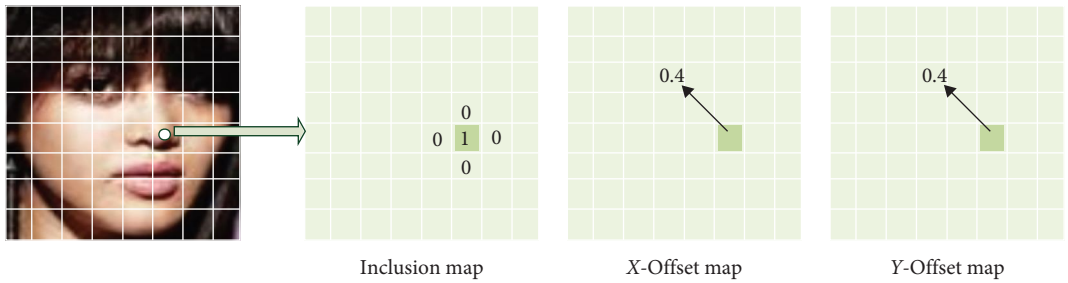

Figure 5: Inclusion map, X-direction offset map and Y-direction offset map obtained by the hybrid detection model.

the continuity and relative position stability of landmarks, which is easily affected by occlusions or large posture movements.

Considering the advantages and disadvantages of the two methods mentioned above, we adopt the detection model combining heatmap regression and coordinate regression [6], as shown in Figure 4, to combine the advantages of heatmap regression and coordinate regression and complement each other. As shown in Figure 4, facial features are extracted through a series of convolution operations through the residual network, and the size of the feature map is reduced to 8×8. Then, we use three convolution operations instead of upsampling to get the inclusion graph and the X and Y-offset graph. Therefore, the output of this mixed regression model is three, which are the inclusion graph ($N \times H \times W$) and the X and Y-offset graph ($2N \times H \times W$), respectively, as shown in Figure 4 below.

More specifically, the image with the input size of 256×256 is downsampled into a feature map with the size of 8×8, which is divided into 64 blocks. Among them, the block on which the key point is located is assigned a value of 1, and the rest are 0. At this time, the feature map obtained predicts the inclusion of the key point. Alternatively, we may denote the left or upward extension of the block as the positive direction of the X axis or the positive direction of the Y axis, respectively, as the origin of the coordinate axis. If the key point is offset by 40% in the X-axis direction relative to the origin, then its X-offset is denoted as 0.4. Then its Y-offset is equal to 0.4, as shown in Figure 5 below.

At the same time, most of the training dataset is composed of frontal face images, so the samples with large poses on the side account for a small proportion, which leads to the problem of data imbalance. Training will focus on the features of frontal samples and ignore the features of the large pose. Inspired by the paper of Guo et al. [28], we add a head attitude prediction module into the backbone network, as shown in Figure 6, to predict the Euler angles of head

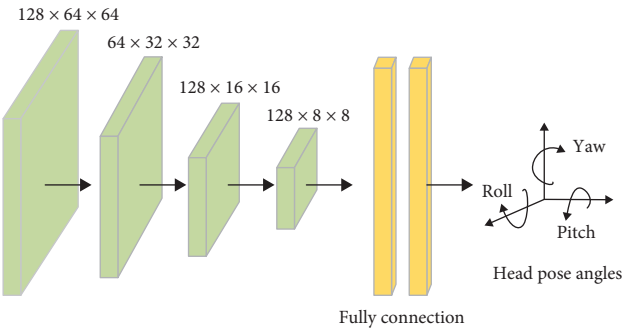

Figure 6: Head pose prediction module.

three-dimensional attitude rotation, namely pitch angle, yaw angle, and roll angle, and obtain the angle information of face rotation. As shown in Figure 6, three Euler angles are obtained through several convolutions and full connection layer regression. Then, the obtained Euler angle is used to add an adaptive weight module to the loss function. When the angle is larger, the penalty will be larger, which makes the model pay more attention to large pose samples with fewer numbers and enhances its robustness. On the contrary, when the angle is smaller, the penalty will be smaller. The head pose prediction module can be completed by several simple convolution operations.

3.3. Offset Prediction Networks. In the previous initial prediction network, we obtained the initial coordinates of the landmarks. However, the coordinates obtained by the mixed detection model based on the block of coordinates predicted by heatmap regression and then the offset prediction are not accurate enough. Our backbone network uses a mixed detection model is based on the heatmap regression. Actually this is a feature matching process, so it still pays more attention to local characteristics. In the initial inclusive classification of key points, key points belonging to a certain block can share

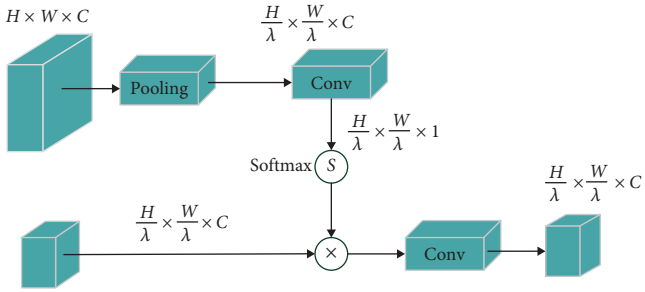

FIGURE 7: Structure of attention fusion module.

the same characteristics, so as to achieve partial continuity. However, for key points belonging to two different blocks, the prediction is based on the different feature blocks, so they are independent of each other. To some extent, this still destroys the continuity of the landmarks of the face, without further considering the spatial relationship between the points. Since, coordinate regression is based on the same features, global continuity is preferred. Therefore, based on the coordinate regression and attention mechanism, we regress a series of offsets of initial coordinates relative to ground truth to further refine the coordinates of landmarks.

We take the generated heatmap corresponding to the features pre-extracted from the network and the initial coordinates of the first stage as input to this offset prediction network. As shown in Figure 2, this network is based on coordinate regression, takes VGG16 [29] as the network baseline, and the generated heatmap as the attention mask to guide the offset prediction network to pay more attention to the area around the key point and apply weights to the extracted features. Figure 7 shows the generation and fusion process of the attention mask. As shown in the figure, the input heatmap generates the attention mask after downsampling and several convolution and softmax operations, and readds the weight for the pre-extracted features, which is conducive to modeling the context relationship between the facial regions. The process of attention mask generation and fusion can be expressed by a mathematical formula as follows:

$$f = S(g(C * D(H) + b)) \otimes F, \tag{1}$$

where C stands for the downsampling operation, g for the nonlinear function, S for the softmax operation, F for the feature extracted in the previous step, and $\otimes$ for the element-by-element dot product operation. This is helpful for the network to more accurately capture the spatial relationship between the segmented regions of the face. To optimize the offset results predicted by the network, we calculate the predicted offset and the mean square error of the offset between the initial coordinates and the ground truth. The objective function of the network can be defined as follows:

$$F = \arg\min \frac{1}{N} \sum_{i=1}^{N} \| S_i - S'_i - O_i \|_2, \tag{2}$$

where N represents the total number of facial landmarks, S_i donates the ground truth, S'_i represents the predicted initial coordinates, and O_i represents the predicted offset. This objective function will force the network to learn the offset between the initial coordinates and the ground truth, effectively combine the facial space relationship, further refine the coordinates of the landmarks, and combine the two-stage network output to obtain the final coordinates.

3.4. Loss Function. Since the regression output of our network is divided into two parts, namely, inclusion map and offset, the loss function consists of the following two parts, as shown in Equation:

$$L = L_{\text{inclusion}} + \alpha L_{\text{offset}}, \tag{3}$$

where L represents the total loss, $L_{\text{inclusion}}$ contains the graph loss, $L_{\text{offset}(x,y)}$ represents the x and y offset loss, and α is the equilibrium coefficient.

For the inclusion map loss, it can be expressed as the follows:

$$L_{\text{inclusion}} = \frac{1}{NHW} \sum_{n=1}^{N} \sum_{i=1}^{H} \sum_{j=1}^{W} \gamma_n \left(E_{ij}^{(n)} - \widehat{E}_{ij}^{(n)} \right)^2, \tag{4}$$

where N is the number of landmarks, H and W are the height and width of the inclusion map. γ_n is the adaptive weight, $E_{ij}^{(n)}$, and $\widehat{E}_{ij}^{(n)}$ are the true and predicted inclusion graphs, respectively. The $E_{ij}^{(n)}$ equals 0 or 1.

For offset map loss, it can be expressed as the following equation:

$$L_{\text{offset}} = \frac{1}{2N} \sum_{E_{ij}^{(n)}=1} \sum_{l=1}^{2} \gamma_n \left| O_{ij}^{(n,l)} - \widehat{O}_{ij}^{(n,l)} \right|, \tag{5}$$

where $O_{ij}^{(n,l)}$ and $\widehat{O}_{ij}^{(n,l)}$ are the true and predicted offset, respectively. l represents the x and y directions. The $O_{ij}^{(n,l)}$ is between 0 and 1.

γ_n in Equations (4) and (5) above is the adaptive weight composed of Euler angles obtained from the head pose prediction module, which is defined as follows:

$$\gamma_n = \sum_{k=1}^{K} \left(1 - \cos\theta_n^k \right), K = 3, \tag{6}$$

here θ_n^k stands for the difference angle between the predicted Euler angles corresponding to each point and the ground truth, respectively. K stands for the number of Euler angles. Obviously, with the increase of Euler angle, the penalty of this term for the whole loss will be larger, which makes the model pay more attention to the large pose face with fewer samples and increases the robustness of the model.

4. Experiments

In this section, we introduce the datasets used in this paper, including WFLW [9], COFW [8], 300W [7], implementation details, and experimental results analysis.

4.1. Datasets. We selected three challenging datasets to test the performance of AWMOP, namely WFLW, COFW, and 300W.

WFLW: the WFLW dataset is annotated with 98 landmarks and contains 10,000 images, of which 7,500 are used for training and 2,500 are used for testing. At the same time, the dataset also annotates the categories of the samples, and there are six categories, which are: pose, expression, lighting, makeup, occlusion, and blur. Therefore, the test set is also divided into six subsets corresponding to the above six categories, and each subset is merged into the complete test set.

COFW: the COFW dataset presents the face state in the real world with a total of 29 landmarks, which is more challenging in the case of occlusion and large poses. We use COFW color images for training and testing. There are 1,345 images in the training set and 507 images in the test set.

300W: 300W dataset is a very general dataset with 68 landmarks. This dataset is divided into four subsets, which are AFW, HELEN, IBUG, and LFPW. A total of 3,148 sample images in IBUG and training subsets of HELEN and LFPW were used as the training set, and 689 sample images in IBUG and testing subsets of HELEN and LFPW were used as the test set. The whole test set is divided into a challenging subset and a common subset. At present, there are two normalization standards for this dataset, namely, inter-pupil normalization and eye spacing normalization.

4.2. Implementation Details. Before training, in order to retain more features and context information, sample images of the 300W dataset were expanded outward by 10% and then trimmed according to the bounding box given by annotation, the WFLW dataset was expanded outward by 20% and then trimmed, and COFW dataset was trimmed directly according to the bounding box given. Resize the cropped image to 256×256 and input it into the network. We performed the random translation of the X-axis and Y-axis with ± 30 pixels, random occlusion of a rectangle with a maximum length of 100 pixels, and horizontal flip operation with probability $P = 0.5$ for the sample data. We used the pretrained network on ImageNet [29] as the backbone network and Adam as the optimizer. The initial learning rate was set to 0.0001. The backbone network was trained with 60 epochs and decayed by 10 each at the 30-th and 50-th epochs. The balance coefficient in the loss function is set to 0.1. The baseline network uses pretrained VGG16 on ImageNet, the initial learning rate is set to 0.00005 and decayed by 5 after 100 epochs. To prevent the overfitting of the backbone network in the first stage, different training subsets are randomly used to train the backbone network and the offset prediction network, respectively, and then trained jointly.

4.3. Metrics. To test the performance of the model, the widely used normalized mean error (NME) and cumulative error distribution (CED) curves were used as the evaluation indexes of the model performance. The mathematical formula of NME is defined as follows:

$$\mathrm{NME} = \frac{1}{M} \sum_{m=1}^{M} \frac{\frac{1}{N} \sum_{n=1}^{N} \|\widehat{p}_{mn} - p_{mn}\|_2}{d_m}, \tag{7}$$

where M is the total number of test images, N is the number of landmarks, $\widehat{p}_{mn}$ and p_{mn} represent the predicted and true coordinates of the n-th key point of the m-th image, respectively. d_m represents the normalized distance. For the 300W dataset, we used two normalization criteria, namely, inter-pupil distance normalization (IPN) and inter-ocular distance normalization (ION). However, for COFW and WFLW datasets, we used the normalization standard of inter-ocular distance.

The mathematical formula of CED is defined as follows:

$$\mathrm{CED} = \frac{N_{e \le l}}{N}, \tag{8}$$

where $N_{e \le l}$ is the number of images which error l no less than e.

4.4. Experiment Results on WFLW. To evaluate the performance of AWMOP under various interference situations, we tested the model using the WFLW dataset. We show the NME results of the model in various situations and compare them with other advanced methods, including LBF [5], ESR [30], CFSS [4], DVLN [31], LAB [9], Wing [32], FCDN [33], PIPNet [6], MAttHG [21], and SD-HRNet [34].

Figure 8 shows the subjective results of our model on the WFLW dataset. Each column from left to right contains six subsets of images: pose, expression, lighting, makeup, occlusion, and blur. It is apparent from the test results that the six subsets produced good results. In addition, one picture may belong to several different subsets. For example, the fifth picture in the first row belongs to both the blur subset and the occlusion subset. Figure 8 shows that when several different interference situations occur at the same time, we can also achieve good results.

In Table 1, we show the NME results of the model on the full test set of WFLW and the other six subsets. We displayed the optimal results in bold. As can be seen from the table, AWMOP has the largest effect and ranks first in the test set. Meanwhile, in the other six subsets, AWMOP also performed best in the expression and makeup subsets. And the performance of the large pose, illumination, and blur subsets ranked second, indicating that the model has high robustness to the interference of large pose, exaggerated expression, and makeup. However, by comparing the test results of each subset, it can be found that the test results of the pose, occlusion, and blur subsets are significantly worse than those of the other subsets, which indicates that there is still a lot of room for improvement for the extreme pose, extreme occlusion, and blur.

4.5. Experiment Results on COFW. In order to evaluate the robustness of AWMOP against occlusion and large attitude, we use COFW to test the model and write the test results in Table 2. As shown in Table 2, we compared the NME and failure rate with other existing advanced methods using eye

Figure 8: Subjective results of the model on WFLW. Each column from left to right contains six subsets of images: pose, expression, lighting, makeup, occlusion, and blur.

Table 1: Comparison of results with the state-of-the-art method on the WFLW dataset, error (NME) normalized by the inter-ocular distance.

Method	Test	Pose	Expression	Illumination	Makeup	Occlusion	Blur
LBF [5]	10.29	24.10	11.45	9.32	9.38	13.03	11.28
ESR [30]	11.13	25.88	11.47	10.49	11.05	13.75	12.20
CFSS [4]	9.07	21.36	10.09	8.30	8.74	11.76	9.96
DVLN [31]	6.08	11.54	6.78	5.73	5.98	7.33	6.88
LAB [9]	5.27	10.24	5.51	5.23	5.15	6.79	6.32
Wing [32]	5.11	8.75	5.36	4.93	5.41	6.37	5.81
FCDN [33]	4.86	**7.81**	5.13	4.77	4.77	5.78	**5.34**
PIPNet [6]	4.79	8.76	4.86	4.56	4.60	6.04	5.49
MAttHG [21]	4.68	8.18	4.83	**4.48**	4.71	**5.26**	5.72
SD-HRNet [34]	4.93	8.63	5.31	4.81	4.76	5.73	5.56
AWMOP	**4.68**	8.18	**4.81**	4.54	**4.41**	5.79	5.38

Bold values represent the best results.

Table 2: Comparison of results with the state-of-the-art methods on the COFW dataset, with mean error (NME) normalized by inter-pupil distance.

Method	Mean error	Failure rate
RCPR [8]	8.76	20.12
LBF [5]	8.77	–
OSRD [35]	9.27	–
TCDCN [11]	7.66	16.17
CFSS [4]	6.28	9.07
LAB [9]	5.58	2.76
Wing [32]	5.44	3.75
FCDN [33]	5.32	2.17
MAttHG [21]	5.08	1.26
SD-HRNet [34]	**3.69**	**0.2**
AWMOP	4.83	0.98

Bold values represent the best results.

spacing as the normalization standard, including RCPR [8], LBF [5], OSRD [35], TCDCN [11], CFSS [4], LBA [9], Wing [32], FCDN [33], MAttHG [21], and SD-HRNet [34]. We have bolded the best result.

From the results in the Table 2, we can see that we have achieved better results in terms of both NME and failure rate. Its predictive effect is only inferior to SD-HRNet's. Due to the more challenging occluded images in the COFW dataset, they contain varying degrees of occlusion in different parts. As shown in the results on the WFLW dataset, our method has a good effect on occluded images to some extent, but it still cannot handle cases of partial or complete occlusion well. So our prediction performance will be slightly worse than SD-HRNet [34], which can maintain high resolution throughout the entire process. Figure 9 shows the cumulative error distribution curves of RCPR, TCDCN, CFSS, 2-SCRM, and SHN-GCN of the proposed method on the COFW

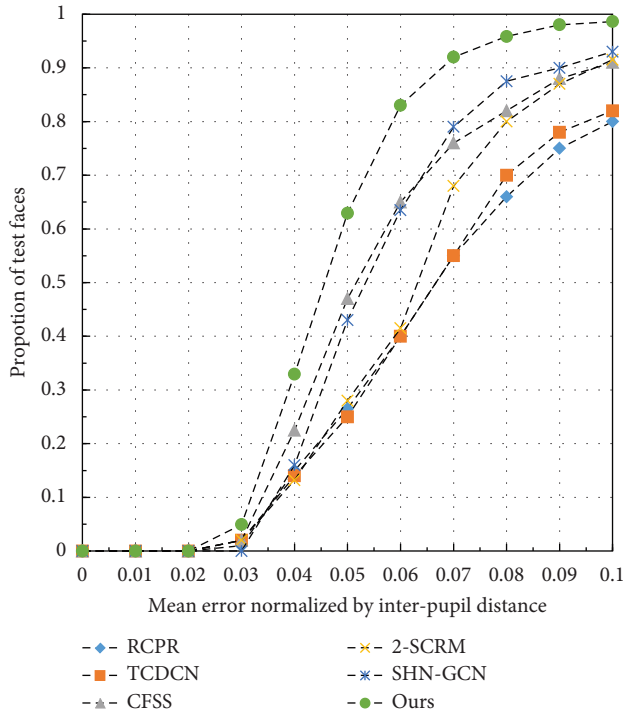

Figure 9: CED curve for testing the proposed method on the COFW dataset.

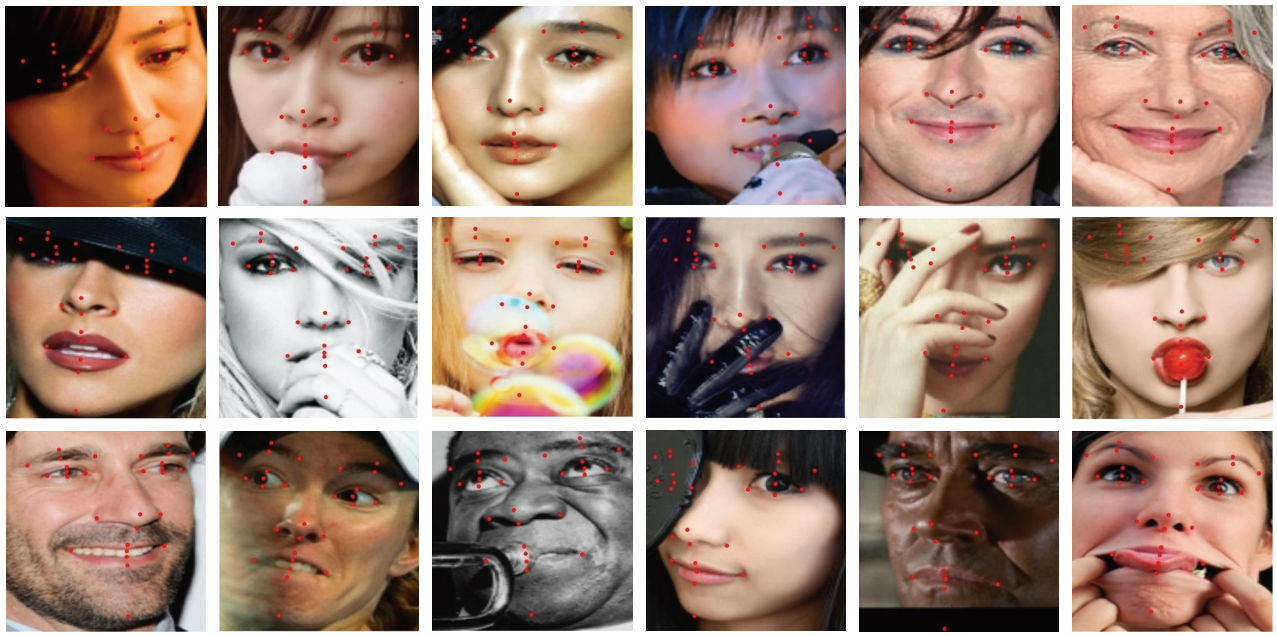

Figure 10: Detection effect of the model on COFW dataset. The first row is the case of slight occlusion, the second row is the case of heavy occlusion, and the third row is the case of both occlusion and large pose or exaggerated expression.

test set. As we all know, the higher the trend of the CED curve, the better the positioning effect. As shown in the figure, the CED curve of our method is higher than that of other methods, which further illustrates the advantages of our method.

Figure 10 shows the excellent performance of AWMOP on occlusion and different poses. The first line in Figure 10 shows the testing effect of the image with slight occlusion, which shows that the positioning effect is accurate. The second line shows the test results of the pictures with the most occlusions. It can be seen that if the occlusions are only partially occluded, the test results are more accurate, such as the eyes in the upper right corner of the second line that is partially occluded by the hair, and the nose and mouth parts in the middle and lower part of the third line that is occluded

Table 3: Comparison of results with state-of-the-art methods on the 300W dataset.

Method	Common	Challenging	Full
Inter-pupil normalization (IPN)			
RCPR [8]	6.18	17.26	8.35
ESR [30]	5.28	17.00	7.58
LBF [5]	5.57	15.40	7.50
TCDCN [11]	4.80	8.60	5.54
CFSS [4]	4.73	9.98	5.76
2-SCRM [15]	4.58	10.94	5.82
MAttHG [21]	**3.83**	**7.12**	**4.49**
AWMOP	4.09	7.63	4.79
Inter-ocular normalization (ION)			
DSRN [16]	4.12	9.68	5.21
DSCN [17]	3.58	5.36	3.85
LAB [9]	2.98	5.19	3.49
MAttHG [21]	**2.81**	**4.94**	**3.23**
SD-HRNet [34]	2.94	5.33	3.41
AWMOP	2.95	5.28	3.41

Bold values represent the best results.

by transparent bubbles, etc. However, if local features are all occluded, the results will be offset, resulting in inaccurate positioning, such as the second line of the sixth left corner of the eye part. The third row shows the test results of images with occlusion and attitude transformation or exaggerated expression. It can be seen that although the same situation exists in the severely occluded parts, the accuracy of the overall shape and local positioning is maintained. Our model works well for detecting large and occluded pose images, while it needs to be refined for localizing fully occluded features.

4.6. Experiment Results on 300W. In order to evaluate the performance of the model on the 300W dataset, we summarized the NME on the full test set and the common and challenge subsets as the test results in Table 3. The other method data in the table comes from their original paper, including RCPR [8], ESR [30], LBF [5], TCDCN [11], CFSS [4], LAB [9], 2 - SCRM [15], SLPT [24], MAttHG [21], DSNR [16], DSCN [17], and SD-HRNet [34]. The best results have been bolded.

As you can see, our method outperforms the vast majority of the current methods in Table 3. It can be seen that our method's objective performance on the common subset of 300W is slightly inferior to the top-ranked MAttHG. When using the distance between pupils as the normalization standard (IPN), our method ranks second. When using the distance between eyes as the normalization standard (ION), our method ranks second alongside SD-HRNet, only slightly lags in the challenge subset. This may be because SD-HRNet can learn the advantages of the original HRNet [36] through knowledge distillation, maintaining high resolution throughout the network and obtaining richer information. Among the results under two normalization criteria, the difference between our method's results on the common subset and MAttHG is smaller than that on the challenge subset. The overall experimental performance is not as good as that of the WFLW dataset. It may be that learning offset requires more sample support. In the case of limited training samples, the position attention module and channel attention module used in the MAttHG method are more able to learn the changes in local detail features caused by large poses and large expressions. Figure 11 shows the visualization results on the 300W dataset, which are, respectively, the public subset and challenging subset. It can be seen that our detection accuracy is good whether it is a frontal face in the public subset or challenge subset or face image with occlusion, large pose, and exaggerated expression. For example, the irregular contours of the first, sixth, and eighth pictures in the first row all fit well, and for the faces with local exaggerated expressions such as the sixth, seventh and eighth pictures in the first and second rows, our detection effect is also good.

Figure 12 shows the comparison between the results of our method on 300W and those of other methods. From left to right, each column is (a) ground truth, (b) 2-SCRM [15], (c) CFSS [4], (d) LBF [5], and (e) the results of our method. From the results of the third, fourth and fifth lines in the figure, we can see that our method has a good positioning effect for these large poses and partially occluded images, and is more accurate in facial contour.

4.7. Ablation Test. Our proposed model consists of three parts, namely, the MFP module, head attitude prediction module, and offset prediction network module. In this section, we will verify their effectiveness on 300W, COFW, and WFLW datasets, respectively. We have bolded the best results in each table.

(1) Analysis of the MFP module. Before the start of the network, we add the MFP module, which can extract multi-scale local feature information, which is very helpful for the subsequent learning of our model. To verify the effectiveness of this component, the results of the model without the pre-extraction module and the model with the pre-extraction module are compared, and the test results are shown in Table 4. From the table, we can see that the NME of each dataset has decreased to a certain extent. It is worth mentioning that, as shown in Table 5, after adding the pre-extraction module, the makeup and occlusion subsets are greatly improved, and the expression subset is also improved to a certain extent, which indicates that the MFP module is effective for the extraction of local features, and it is beneficial for makeup, exaggerated expression, occlusion, and other situations.

(2) Analysis of the head pose prediction module. We add the head pose prediction module to the backbone network to predict the rotation angle information of the head and give more weight to the large pose images in the dataset so that the network can learn features better. In order to verify its effectiveness, we conducted experiments based on the original backbone network and compared the results with the head pose prediction module. The experimental results are shown in Table 6. Compared to the original backbone

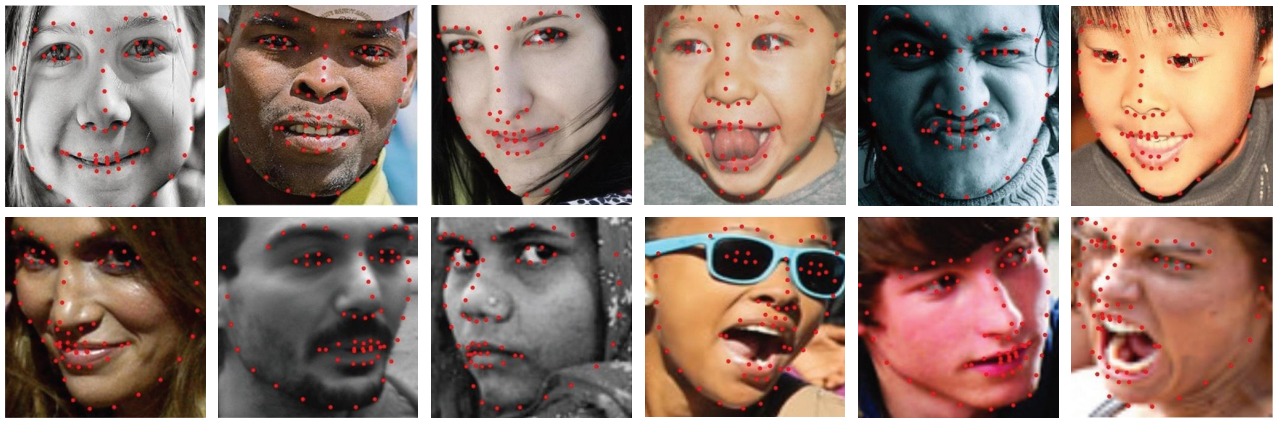

Figure 11: 300W dataset subjective visualization display. The first row is the common subset, and the second row is the challenge subset.

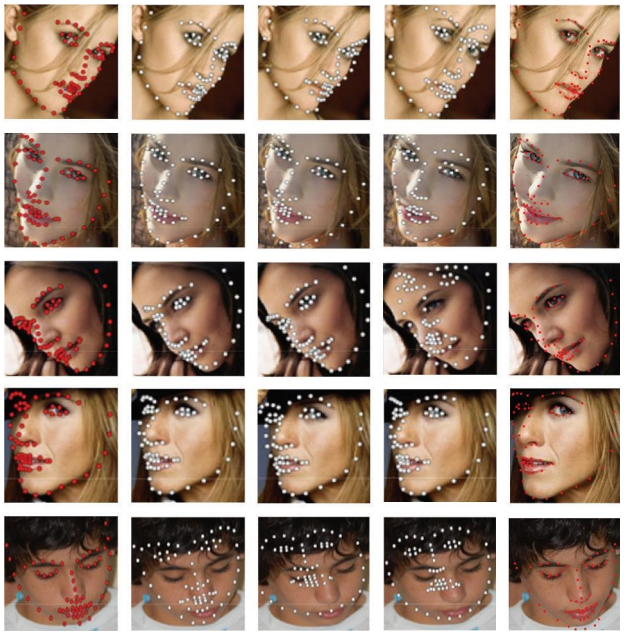

Figure 12: The comparison between the results of our method on 300W and those of other methods. From left to right, each column is (a) ground truth, (b) 2-SCRM, (c) CFSS, (d) LBF, and (e) the results of our method.

Table 4: Comparison of mean errors with or without multi-scale feature pre-extraction (MFP) modules on the 300W test complete set, COFW, and WFLW datasets.

Settings	300W	COFW	WFLW
Without MFP	4.83	4.85	4.70
With MFP	**4.79**	**4.83**	**4.68**

Bold values represent the best results.

network, our addition of head pose prediction provides better results. At the same time, we can see in Table 5 that the NME of the model with this component on the large pose subset decreases by 0.18%, which indicates that the addition of the head pose prediction module is indeed very beneficial to the large pose sample and also improves the overall effect.

(3) Analysis of the offset prediction network module. As different blocks depend on the different features during heatmap regression, some adjacent feature points are discontinuous, so in order to compensate for the global loss caused by heatmap regression, we add an offset prediction network. In order to explore the effectiveness of this network, we compared the test results of the original network and the added offset prediction network. As shown in Table 7, the NME on the three datasets has been improved to some extent. In Table 5, the NME of the pose subset decreases by 0.06%, indicating that the stability of the pose is improved by this module.

(4) Model complexity analysis of components: The model complexity of our components is shown in Table 8.

Table 5: The influence of each component on the detection effect of subsets such as large pose, expression, and makeup.

Settings	Full	Pose	Expression	Illumination	Makeup	Occlusion	Blur
Mixed model	4.74	8.45	4.85	4.63	4.51	5.92	5.49
+ OPN	4.73	8.39	4.84	4.58	4.48	5.92	5.46
+ HPP	4.70	8.21	4.83	4.58	4.48	5.85	5.38
+ MFP	**4.68**	**8.18**	**4.81**	**4.54**	**4.41**	**5.79**	**5.38**

Bold values represent the best results.

Table 6: Comparison of mean errors with or without the head poses prediction (HPP) module on the 300W test complete set, COFW, and WFLW datasets.

Settings	300W	COFW	WFLW
Without HPP	4.84	4.89	4.73
With HPP	**4.83**	**4.85**	**4.70**

Bold values represent the best results.

Table 7: Comparison of the mean errors of the offset prediction network (OPN) with or without the 300W test full set, COFW, and WFLW datasets.

Settings	300W	COFW	WFLW
Without OPN	4.86	4.93	4.74
With OPN	**4.84**	**4.89**	**4.73**

Bold values represent the best results.

Table 8: Model complexity of each component.

Settings	Params	FPS	Average inference time (for each image)
Original	12.33 M	132	0.0081
+ MFP	12.37 M	84	0.0117
+ HPP	12.9 M	83	0.0121
+ OPN	12.9 M + 139.35	15	0.0685

From the table, we can see that the MFP and HPP modules have little impact on the overall model complexity. The fluctuations in the number of model parameters and inference speed per image are acceptable when compared with the improvements in model detection performance. However, the OPN module uses a more complex framework, VGG16, to implement the auxiliary network and therefore increases inference speed. How to lighten the model is still a problem that requires continued research.

In summary, we can see the effectiveness of our various components in processing facial images in different situations. We add a MFP module at the early stage of the network to enable the subsequent network to obtain features that integrate information from various scales. The parallel connection of high-resolution and low-resolution feature maps preserves much richer local information, thereby reducing the NME of the network in makeup and occlusion test sets by 0.07% and 0.06%. The head pose prediction module balances the data by applying greater weights to the large pose images in the training set, enhancing the model's prediction effect on large pose images. As shown in Table 5, the addition of this module reduces the NME of the large pose test subset by 0.18%. Finally, the offset prediction network further optimized the coordinates, and the overall model's performance in the large pose test subset was further improved, with NME decreasing by 0.06% again. Overall, due to the addition of MFP and head pose prediction components in our model, it is more suitable for dealing with the impact of large poses, occlusion, and heavy makeup on images in natural environments, improving the robustness of the model to such complex situations. At the same time, with the assistance of the offset prediction module, the model is enhanced to constrain the overall shape, further improving the prediction accuracy of large pose faces.

5. Conclusion

In this paper, an adaptive weighted face AWMOP is proposed to solve the problems of a large pose, exaggerated expression, and excessive makeup in the unconstrained field environment. First, a MFP module is added to obtain multiscale features and focus on local features. Then, the head poses prediction module is introduced into the backbone network to increase the weight of large pose images to solve the problem of data imbalance. Meanwhile, an offset prediction network is added to improve the spatial relationship between landmarks and optimize the global shape by relying on global features. Good results are obtained on common public challenge sets: 300W, COFW, and WFLW, which shows the superiority of our model.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This study is supported by National Natural Science Foundation of China, No.12201075.

References

[1] M. Hao, G. Liu, and D. Xie, "Hyperspectral face recognition with a spatial information fusion for local dynamic texture patterns and collaborative representation classifier," *IET Image Processing*, vol. 15, no. 8, pp. 1617–1628, 2021.

[2] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3D face model from in-the-wild images," *ACM Transactions on Graphics*, vol. 40, no. 4, pp. 1–13, 2021.

[3] P. R. S. Devi and R. Baskaran, "SL2E-AFRE: Personalized 3D face reconstruction using autoencoder with simultaneous subspace learning and landmark estimation," *Applied Intelligence*, vol. 51, no. 4, pp. 2253–2268, 2021.

[4] S. Zhu, C. Li, C. Change Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4998–5006, IEEE, 2015.

[5] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1685–1692, IEEE, 2014.

[6] H. Jin, S. Liao, and L. Shao, "Pixel-in-pixel net: towards efficient facial landmark detection in the wild," *International Journal of Computer Vision*, vol. 129, no. 12, pp. 3174–3194, 2021.

[7] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: database and results," *Image and Vision Computing*, vol. 47, pp. 3–18, 2016.

[8] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1513–1520, IEEE, 2013.

[9] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, "Look at boundary: a boundary-aware face alignment algorithm," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2129–2138, IEEE, 2018.

[10] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 386–391, IEEE, 2013.

[11] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., vol. 8694 of *Lecture Notes in Computer Science*, pp. 94–108, Springer, Cham, 2014.

[12] Y. Wu, T. Hassner, K. G. Kim, G. Medioni, and P. Natarajan, "Facial landmark detection with tweaked convolutional neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 3067–3074, 2018.

[13] W. Li, Y. Lu, K. Zheng et al., "Structured landmark detection via topology-adapting deep graph learning," in *European Conference on Computer Vision*, pp. 266–283, Springer, 2020.

[14] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3d dense face alignment," in *European Conference on Computer Vision*, pp. 152–168, Springer, 2020.

[15] J. Zhang, L. Di, and J. Liang, "Face alignment based on fusion subspace and 3D fitting," *IET Image Processing*, vol. 15, no. 1, pp. 16–27, 2021.

[16] R. Valle, J. M. Buenaposada, A. Valdes, and L. Baumela, "A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 585–601, ECCV, 2018.

[17] A. Ullah, J. Wang, M. S. Anwar, T. K. Whangbo, and Y. Zhu, "Empirical investigation of multimodal sensors in novel deep facial expression recognition in-the-wild," *Journal of Sensors*, vol. 2021, Article ID 8893661, 13 pages, 2021.

[18] J. Wan, Z. Lai, J. Li, J. Zhou, and C. Gao, "Robust facial landmark detection by multiorder multiconstraint deep networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 5, pp. 2181–2194, 2022.

[19] Y. Huang, H. Yang, C. Li, J. Kim, and F. Wei, "Adnet: Leveraging error-bias towards normal direction in face alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3080–3090, IEEE/CVF, 2021.

[20] A. Bulat and G. Tzimiropoulos, "Convolutional aggregation of local evidence for large pose face alignment".

[21] Z. Yang, X. Shao, J. Wan, R. Gao, and Z. Lai, "Mixed attention hourglass network for robust face alignment," *International Journal of Machine Learning and Cybernetics*, vol. 13, no. 4, pp. 869–881, 2022.

[22] J. Xie, J. Wan, L. Shen, and Z. Lai, "Think about boundary: fusing multi-level boundary information for landmark heatmap regression," in *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2021.

[23] H. Park and D. Kim, "ACN: Occlusion-tolerant face alignment by attentional combination of heterogeneous regression networks," *Pattern Recognition*, vol. 114, Article ID 107761, 2021.

[24] J. Xia, W. Qu, W. Huang, J. Zhang, X. Wang, and M. Xu, "Sparse local patch transformer for robust face alignment and landmarks inherent relation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4052–4061, IEEE, 2022.

[25] H. Li, Z. Guo, S.-M. Rhee, S. Han, and J.-J. Han, "Towards accurate facial landmark detection via cascaded transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4176–4185, IEEE, 2022.

[26] C. Tang, X. Liu, S. An, and P. Wang, "Br^{2}net: defocus blur detection via a bidirectional channel attention residual refining network," *IEEE Transactions on Multimedia*, vol. 23, pp. 624–635, 2021.

[27] C. Tang, X. Liu, X. Zheng et al., "DeFusionNET: defocus blur detection via recurrently fusing and refining discriminative multi-scale deep features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 955–968, 2022.

[28] X. Guo, S. Li, J. Yu et al., "Pfld: a practical facial landmark detector," 2019.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.

[30] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 177–190, 2014.

[31] J. Zhang, M. Kan, S. Shan, and X. Chen, "Leveraging datasets with varying annotations for face alignment via deep regression network," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3801–3809, IEEE, 2015.

[32] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2235–2245, IEEE, 2018.

[33] J. Wan, Z. Lai, L. Shen et al., "Robust facial landmark detection by cross-order cross-semantic deep network," *Neural Networks*, vol. 136, pp. 233–243, 2021.

[34] X. Lin, H. Zheng, P. Zhao, and Y. Liang, "Sd-hrnet: slimming and distilling high-resolution network for efficient face alignment," *Sensors*, vol. 23, no. 3, Article ID 1532, 2023.

[35] M. Rogers and J. Graham, "Robust active shape model search," in *European Conference on Computer Vision*, pp. 517–530, Springer, 2002.
[36] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5693–5703, IEEE, 2019.