Cognitive Computation (2024) 16:1964–1978
https://doi.org/10.1007/s12559-022-10065-9

Deep Recurrent Regression with a Heatmap Coupling Module for Facial Landmarks Detection

M. Hassaballah[1,2] · Eman Salem[3] · Abdel-Magid M. Ali[3] · Mountasser M. Mahmoud[3]

Received: 7 April 2022 / Accepted: 5 October 2022 / Published online: 27 October 2022

Abstract

Facial landmarks detection is an essential step in many face analysis applications for ambient understanding (people, scenes) and for dynamically adapting the interaction with humans and environment. The current methods have difficulties with real-world images. This paper proposes a simple and effective method to detect the essential points in human faces. The proposed method comprises a two-stage coordinated regression deep convolutional neural network (CR-CNN) with a heatmap coupling module to convert the detected facial landmarks of the first stage into a Gaussian heatmap. To take advantage of the prior stage knowledge, the generated heatmap is concatenated with the original image of the input face and entered into the network in the second stage. The two-stage implementation based on CR-CNN has same layers structure to simplify the design and complexity. The L_1 loss function is used for each stage and the total loss equals the sum of the two loss functions from both stages. Comprehensive experiments are conducted to evaluate the proposed method on three common challenging facial landmark datasets, namely AFLW, 300W, and WFLW. The proposed method achieves normalized mean error (NME) of 1.56% on the AFLW, 4.20% on the 300W, and 5.53% on the WFLW datasets. Moreover, the execution time of the proposed two-stage CR-HC is calculated as 3.33 ms. The obtained results show the robustness and outstanding performance of the proposed method over some of the state-of-the-art methods. The source code is provided as an open repository to the community for further research activities.

Keywords Face analysis · Facial landmarks detection · Pose estimation · Deep learning · Coordinates regression · Heatmaps regression

Introduction

In recent years[1], intelligent surveillance systems have been widely studied [1, 2]. The combination of robotics and artificial intelligence arose outstanding developments in the fields of cognitive robotics and human-robot interaction [3]. Nowadays, several academic and industrial research groups are engaged in the design of intelligent robots able to act autonomously using deep learning-based algorithms [4, 5] for the analysis of data acquired from heterogeneous sensors, such as camera, 3D camera, stereo camera, microphone, and LIDAR [6], for ambient understanding (scenes, objects, people) and for dynamically adapting the interaction with humans and environment [7, 8]. Object detection or recognition is one of the most fundamental and challenging problem in computer vision [9, 10]. As a longstanding, challenging problem in object detection, facial landmarks detection (FLD) has been an active area of research for several decades [11]

Facial landmarks detection, also known as face alignment, is the process of locating a specified unique key-point such as the eyes corner, mouth, brows, and tip of the nose [12]. As it is used as a prerequisite for other computer vision applications, detection of these facial points must be robust and reliable. For example, the facial landmarks localization

✉ M. Hassaballah
mah.ali@psau.edu.sa

Eman Salem
eman.salem@aswu.edu.eg

[1] College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, AlKharj, Saudi Arabia

[2] Department of Computer Science, Faculty of Computers and Information, South Valley University, Qena, Egypt

[3] Department of Electrical Engineering, Faculty of Engineering, Aswan University, Aswan 81542, Egypt

[1] The source code is provided as an open repository to the community for further research activities at https://github.com/Eman-salem/CR-HC

are required for many applications like head pose [13], face recognition [14–16], face emotion recognition [17, 18], gender recognition [19, 20], facial beautification [21], as well as facial expression recognition [22]. To ensure the success of these applications, extremely accurate and exceptional detection accuracy is a must. Due to the practical relevance of FLD, the efforts of both industry and academics have been attracted, which in recent years led to significant development. Although the findings have been accomplished, the exact location of facial points in uncontrolled settings remains an exceedingly difficult issue [23, 24]. Besides, a large number of the existing methods are designed based on capturing the local spatial relationship among sets of facial points ignoring that these spatial relationships are high order and global [25].

Cascaded regression is regarded as one of the potential state-of-the-art approaches for refining the prediction of the related predecessor, but the loss of information during the cascading stages makes it fall in complicated cases in the real world [26]. The cascaded deep convolutional neural nets are able to learn a large number of essential filters and combine them in a hierarchical manner to describe latent concepts for features discrimination efficiently, they can withstand high deformations in a human face and extreme pose changes. Considering these capabilities of the cascaded deep convolutional neural nets, they can successfully detect facial landmarks. On the other hand, the loss of spatial information due to resolution, as well as the difficulty of imposing a proper facial form on the collection of estimated landmarks, reduces its accuracy [27]. To solve this issue, we propose using heatmap coupling to prevent the loss of crucial feature information related to the input and transmit this feature to the cascaded layers, where it can be used as variable initialization for the cascaded CNN regressors.

In this regard, the shape S can be progressively refined through estimating the incremental in the shape ΔS, which is needed to be learned within the stage-by-stage methodology [28]. By providing the facial image I and the initial face shape S^0 or even the previous face shape S^{t-1}, the regressor R^t can compute ΔS^t using the image features at each t stage. The main aim of the cascaded regression is to produce the sequence of updates (ΔS^0, ..., ΔS^{t-1}) starting from the initial shape S^0 and converges to S^* (i.e., $S^0 + \sum_{t=0}^{T-1} \Delta S^t \approx S^*$). Based on that the new face shape S^t is updated in a cascade way using

$$S^t = S^{t-1} + R^t\left(I, S^{t-1}\right) \tag{1}$$

where $t = 1, \ldots, T$ and R^t is a linear regressor that can be formulated by

$$R^t = \arg \quad \min_{R^t} \sum_{i=1}^{N} \|(S_i^* - S_i^{t-1}) - R^t(\phi(I_i, S_i^{t-1}))\| \tag{2}$$

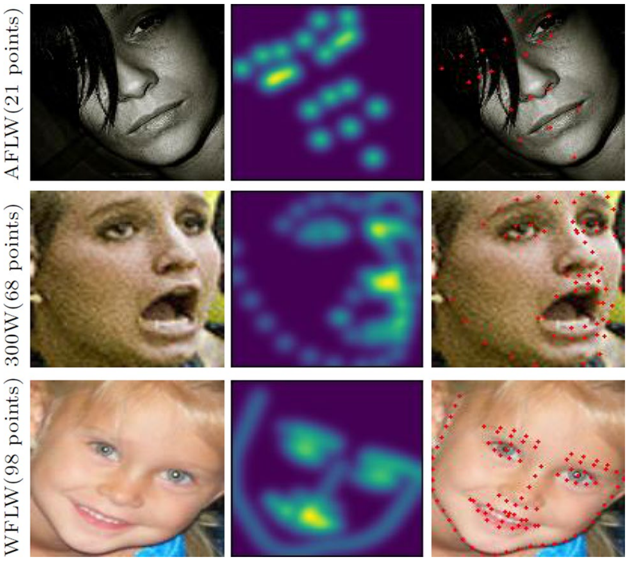

Fig. 1 The first column shows the face images from the datasets with different landmarks annotation. The second column is the output of the heatmap conversion module. The final estimate of the landmarks is shown in the third column

where t refers to the current iteration and R^t is employed to map the feature of the shape indexed $\phi(I_i, S_i^{t-1})$ to the shape residual $(S_i^* - S_i^{t-1})$ and M is the samples number of the training images.

To overcome the drawbacks and limitations of the existing techniques, in this work, we present an accurate and efficient FLD detection method based on a two-stage coordinate regression that is coupled with a heatmap module. The proposed method is called coordinate regression with heatmap coupling (CR-HC). The regression model attempts to extract the shape of the facial landmark as a coarse-to-fine coordinating vector. The input to the first stage is regressed using simple CNN and generates a number of N landmarks. The generated landmarks are transformed using the heatmap module to a Gaussian heatmap with the same dimension as the input image. The second stage is employed to refine the first estimation, which regresses the combination of the input and heatmap images. Figure 1 shows the face images, outputs of the heatmap modules, and the final landmarks estimation with different annotation schemes.

In brief, the main contributions of the work can be summarized as follows.

1. Design a robust deep convolutional neural network model from scratch for facial landmarks detection.
2. Unlike conventional methods based on cascade coordinate regression, we propose a new stage coupling scheme based on a heatmap module to benefit from the input feature for the next stage, which reduces the network complexity.

3. The proposed network is adaptable to be applied on different resolutions images and can achieve comparative results with 128 × 128 resolution despite that it is hard to discriminate the key facial points in case of low-resolution images.
4. Experiments on three challenging benchmark datasets are conducted to evaluate the proposed method, which achieves top performance results in the three datasets compared to state-of-the-art methods.
5. The proposed method's execution time is better for reliable applications than other FLD methods with low execution time but high normalized mean error.
6. Finally, we provide an open repository of the source code to the community for further research activities.

The rest of the paper is organized as follows. The "Related Works'' section introduces a brief discussion about the FLD methods in the literature. The "The Proposed Method'' section describes the proposed FLD method. The evaluated datasets and experiments are presented in the "Experiments and Results'' section. The ablation study was conducted to evaluate the effectiveness of the proposed two-stage CR-HC in the "Ablation Study'' section. Finally, the conclusions and future works are given in the "Conclusion'' section.

Related Works

Facial landmarks detection has made large strides in the last two decades, thanks to technological advancements. It is important that the FLD be more resistant to the static and non-static face deformations caused by occlusion, facial expression, and head motions, notwithstanding the positive results gained [29]. The FLD is nevertheless affected by these conditions, making it unreliable in real-world settings. Generally, the traditional landmark detection methods can be broken down into template-based approaches and regression-based methods [30]. In recent years, deep learning models such as convolution neural networks achieve an enhancement in facial landmarks detection [31], and they can be categorized into coordinate regression and heatmap regression models. The following part presents a brief review of the state-of-the-art methods in the field of facial landmarks detection.

Conventional FLD Approaches

Template fitting models depend on generating a parametric shape from the training dataset and fitting the testing image to this shape during the testing phase. The most popular template-based method is the active shape model (ASM) [32], in which the face shape is represented by a linear combination of fundamental shapes that are learned so that it can use the principal component analysis (PCA). The output shape of the linear model of the shape description S of an object can be formulated as follows

$$S = \bar{S} + \sum_{i=1}^{n} w_i \tilde{S}_i, \tag{3}$$

where $\bar{S}$ is an average example of the object described, w_i is the weighting factor of the model, and $\tilde{S}$ is the i-th object mode. The model is based on the pre-aligned point cloud $\tilde{S}_{1\ldots m}$ from the training set, each sample m from the training set represents a point cloud that describes the shape of an object. The average variable $\bar{S}$ is the average point cloud in $\tilde{S}_{1\ldots m}$, and the model $\tilde{S}_{1\ldots n}$ is the result of the PCA. In addition, the PCA can describe the variation in the appearance of the face shape. The appearance of ASM is modeled by a variety of pre-trained template models, which are the active appearance model (AAM) [33] and PCA models. The appearance in a regular coordinate system eliminates shape alterations and the shape representation is identical to that of ASM and AAM. In [34], a matching approach was proposed for generating a collection of area template detectors using a combined shape and texture appearance model. Despite that these traditional approaches give good results in constrained condition, they are failed on the wild condition as these methods are sensitive to large head pose and occlusion problems. Moreover, both AAM and ASM can not handle the nonlinearity in faces with large head poses as these methods are considered linear in nature; in addition, the irregularity of face shape can lead to self occlusion.

Approaches based on regression immediately learn the mapping from the image to landmarks. It can be a direct regression that can predict the location of the landmark directly without any initialization or cascade regression, which locates the landmarks in a cascade manner depending on the initial shape estimation. The structure information and shape constraints can be learned during the prediction process. The loss function L_2 is usually adopted to calculate the difference between the predicted (S_k) and ground truth (S_k^*) landmarks in a point-wise way as

$$L_2 = \frac{1}{K} \sum_{k=1}^{K} \|S_k - S_k^*\|_2 \tag{4}$$

where K is the number of landmarks set.

For sequential faces, a discriminative response map fitting (DRMF) has been proposed using discriminative regression to estimate model parameters depending on the part-based model in [35]. In [36], the regression forest has been used to estimate the face shape depending on helper facial characteristics such as head pose, gender, etc. [37] proposed an ensemble of regression trees, in which a gradient boosting algorithm is employed to learn each regressor, and it is added to the trees in a cascade manner. Authors of [38] proposed a cascade regression method that utilizes the $L_{2,1}$ normalization factor instead of the least-squares regressor, and multi-initialization is required to increase the regressor robustness for the poor initialization case. In the supervised descent method (SDM), a SIFT of features extracted around the present landmarks is employed to solve a sequence of linear least-squares problems iteratively [39]. A local binary feature is also used to learn a set of local binary features for a cascade regression, as local binary extraction and regressing features are relatively inexpensive computationally [40]. In fact, cascade regression can improve the final facial landmark locations, but it depends on the accuracy of the initial estimation [30]. However, these traditional approaches depend on the handcrafted feature extraction so that some important information in the image is lost and in turn leads to low efficiency in the detection.

Deep Learning-Based Approaches

This type of facial landmarks detection directly maps the face image into the landmark coordinates using deep learning models [41]. In the early work [42], a cascaded CNN is proposed in which the face image is divided into different parts and each part is processed individually using separate deep CNNs. Then, the outputs from each CNN are combined and entered into the final deep CNN to generate final facial coordinates. A Task-constrained deep convolutional network (TCDCN) is proposed for simultaneously optimizing facial landmarks detection with correlated auxiliary tasks such as head pose, gender, and expression [43]. Inspired by knowledge distillation, [25] suggested a loss function for training a lightweight model consisting of two networks, which are the backbone network to regress the coordinates of the facial landmarks and an auxiliary network to estimate the Euler angles of roll, pitch, and yaw. It is worth mentioning that the latter network is used only during the training phase to make the model more practical from the point of model size and processing time. One of the drawbacks of previous methods is that they need special annotated dataset with landmarks and other task annotation to train the model, which is not supported in most of FLD datasets. A recurrent neural network and deep neural network are utilized to estimate the facial coordinates in [44]. This model consists of two networks, a global network with long short-term memory to estimate the initial shape, while the other network utilizes a component-based search method to generate the final shape. In [45], a two-stage branched convolutional neural network (BCNN-JDR) combined with Jacobian deep regression was proposed. The initialization consists of a branched CNN to estimate the face parts individually and the refinement stage to refine the result in a cascade manner. The work of [46] pays more attention to the loss function that is used to train the facial landmark detection model by designing a new loss function named rectified wing loss (Rwing). The developed loss function can handle small-medium error in a good manner compared to the conventional loss function. Although the coordinate regression is simple and fast, but it is not accurate and needs to be handled in a cascade manner to give high accuracy and this sometimes leads to the loss of information during the cascading.

Heatmap regression is the process of finding the likelihood of specific key points residing in the ground truth heatmaps. This type of method usually uses the fully convolutional framework so that it can regress multiple heatmaps keeping the same size as the input image. To address the facial landmarks detection problem, [47] presented a multi-order multi-constraint deep network (MMDN) based on the consolidation of an implicit multi-order correlated geometry aware model and the explicit probability-based boundary-adaptive regression (EPBR) method. Moreover, authors in [48] proposed a style aggregated network (SAN) by generating a new styles training dataset with the help of generative adversarial module and then using the generated data with the original to train a heatmap regression network. In [49], a heatmap regression network is proposed based on the strong stacked hourglass network by stacking four of them and improving the stacked hourglass network with hierarchical, parallel, and multiscale residual blocks. Yin et al. [50] try to solve the problem of **2D** heatmap regression complexity by designing an attentive **1D** heatmap regression model through generating two groups of **1D** heatmaps to represent the marginal distributions of x and y coordinates. The real and fake localization are discriminated by using the geometric priors on the face landmarks based on the conditional generative adversarial network (CGAN). The CNN-based face localization is introduced using a coarse and robust heatmap estimation followed by a subsequent regression-based refinement [51]. In such method, there are two sub-networks, the first one tries to estimate the heatmap-based encodings of the location of the facial landmarks. The second subnetwork receives the outputs of the heatmap estimation unit as inputs and refines them by applying the regression. Despite that the heatmap regression provides a good accuracy, it suffers from the complexity, high execution time, and sensitivity to outliers.

The Proposed Method

In this work, we introduce a new facial landmark detection method called CR-HC based on a two-stage coordinate regression model with a heatmap coupling. The proposed method aims to predict N points represented by a shape vector S, where

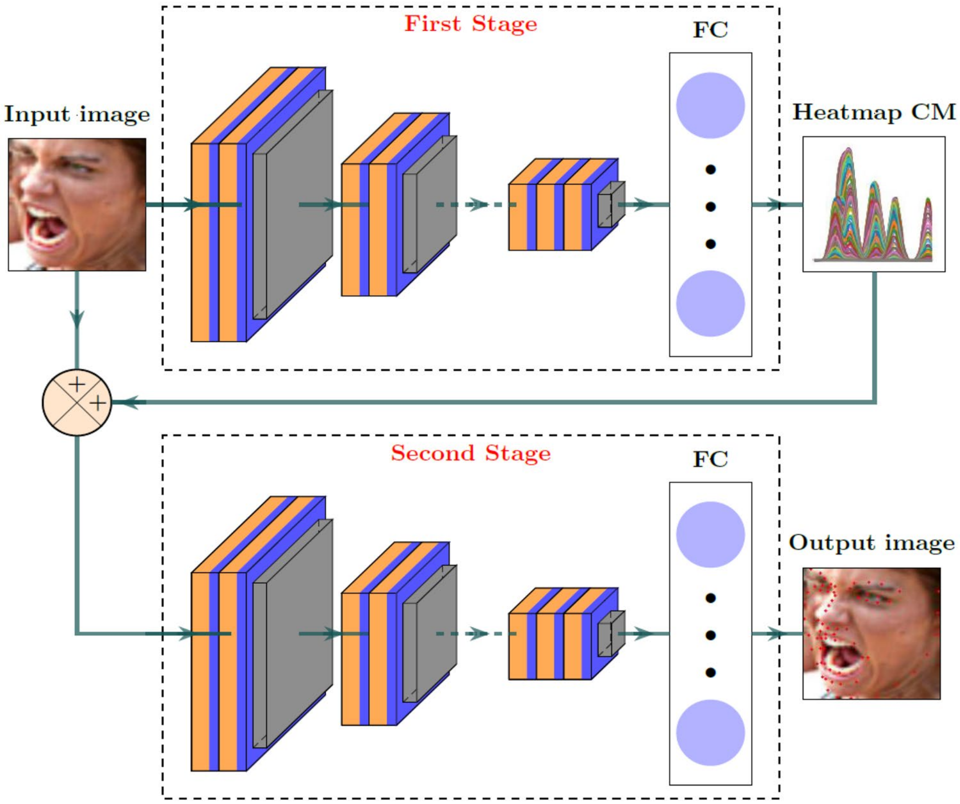

Fig. 2 Structure of the proposed method based on two-stage of coordinate regression neural networks with a heatmap coupling module

$$S = [x_0, y_0, x_1, y_1, \dots, x_n, y_n] = [P_0, P_1, \dots, P_N] \tag{5}$$

where $P_n = (x_n, y_n)$ represents the n_{th} landmark in the face images $I \in \mathbb{R}^{h \times w \times c}$, where h and w are the height and width of the face image, respectively, while c denotes the color channels (e.g., for RGB image, $c = 3$).

The CR-HC method consists of a base regression model and a heatmap coupling module. The regression model aims to extract the shape of the facial landmark as a coordinating vector in a coarse-to-fine manner by stacking the base model to refine the output results with the strong use of the heatmap coupling module. The overall architecture of the proposed model is shown in Fig. 2. A detailed description for each part of the proposed method is discussed in the following subsections.

The Based Model Structure

The backbone network in the proposed model is a custom-built convolution neural network. The design of CNN is intended to be simple and effective to provide flexibility when layered in a multi-stage architecture. It is better to mention here that the number of layers in the proposed two-stage CR-HC is determined by trying a lot of layer configuration and hyperparameters values and choosing those that have the best results. It is made up of stacked convolution blocks, each of them is built with a 3 × 3 convolution layer, followed by batch normalization, and activated with the Relu function. Each stage has seven convolutional blocks, a 2 × 2 pooling layer, and two fully connected (FC) layers.

For a **2***D* image, the convolution operation can be expressed as in (6) in which $k(x, y)$ is the function of each kernel.

$$(I \times k)(x, y) = \sum_{u,v} I(x, y) \times k(x - u, y - v) \tag{6}$$

In the fully connected layer, the input and output images have the same size to reduce the matrix-vector multiplication, while the pooling layer is employed to acquire the invariance against image deformation. It divides the input image into $b \times b$ blocks and chooses the maximum value of each block such that

$$\mathrm{pool}_b(I_{h \times w \times c}) = \max_{0 \le x < b, 0 \le y < b} I_{(h \times b + x) \times (w \times b + y) \times c}. \tag{7}$$

The size of the output feature map is defined according to the number of stride s and padding p of each layer as

$$\begin{aligned} h_{l+1} &= \frac{h_l - h_l' + p}{s} + 1 \\ w_{l+1} &= \frac{w_l - w_l' + p}{s} + 1 \\ c_{l+1} &= m_l \end{aligned} \tag{8}$$

Table 1 Structure of each stage in the proposed method

Input size	Operation	No. ch.	S
128×128×3	Conv,BatchNorm,Relu	64	1
128×128×64	Conv,BatchNorm,Relu	64	1
128×128×64	Max Pooling	64	-
64×64×64	Conv,BatchNorm,Relu	128	1
64×64×128	Conv,BatchNorm,Relu	128	1
64×64×128	Max Pooling	128	-
32×32×128	Conv,BatchNorm,Relu	256	1
32×32×256	Conv,BatchNorm,Relu	256	1
32×32×256	Max Pooling	256	-
16×16×256	Conv,BatchNorm,Relu	512	1
16×16×512	Conv,BatchNorm,Relu	512	1
16×16×512	Max Pooling	512	-
8×8×512	Conv,BatchNorm,Relu	1024	1
8×8×1024	Conv,BatchNorm,Relu	1024	1
8×8×1024	Max Pooling	1024	-
4×4×1024	Conv,BatchNorm,Relu	512	1
4×4×512	Conv,BatchNorm,Relu	512	1
4×4×512	Max Pooling	512	-
2×2×512	Conv,BatchNorm,Relu	256	1
2×2×256	Conv,BatchNorm,Relu	256	1
2×2×256	Full Connection	1024	-
1024	Full Connection	136	-

where l is the number of layer, m denotes the number of kernel unit in a layer l, $\acute{h}$ and $\acute{w}$ are the height and width of the layer's kernel, respectively.

As the proposed method uses the deep convolution model, training such a model can be difficult because they are sensitive to the initial random weights and learning algorithm configuration. This issue is solved by using the batch normalization, which standardizes the inputs to a layer for each mini-batch and reduces generalization error. Table 1 describes the CNN layers in detail.

Let $\mathrm{I}_1 \in \mathbb{R}^{\mathrm{h} \times \mathrm{w} \times \mathrm{c}}$ be the input face to the first stage with 3 color channels (e.g., $c = 3$), where $h \times w$ equals 128×128. The number of channels c in the first convolution block is 64, and the number of channels is doubled in each convolution block, but it is halved in the last two blocks in each stage, as shown in Table 1. The output shape vector of the first stage is $S \in \mathbb{R}^{2 \times N}$, where N is the number of detected landmarks. The output S coordinates vector is converted into a **3***D* heatmap $\mathbf{H} \in \mathbb{R}^{\mathrm{h} \times \mathrm{w}}$ by using the heatmap coupling module. Then, the generated heatmap from the first stage is concatenated with the input face image to be the new features map $\mathrm{I}_2 \in \mathbb{R}^{\mathrm{h} \times \mathrm{w} \times 4}$ that will enter to the second stage. It is noteworthy that the two stages are identical in their structure, but they have different input and output characteristics. Moreover, the coupling point is not the last layer of each stage, but it is approximately located at half of each stage.

The Heatmap Coupling Module

When cascading more levels and making the model deeper, the cascading deep convolution network has lately demonstrated remarkable results in FLD tasks. On the other hand, it suffers from several issues, such as When the processed images are obtained under unconstrained circumstances. Two variables reduce the accuracy of the cascaded model, first, the loss of spatial information reduces the resolution of feature maps in the concatenation of multiple convolutions and pooling layers. In addition, there is an initialization problem, in which the refining process depends on the starting face shape. By providing information to the cascaded stage, the heatmap coupling module is able to resolve the first issue and serve as an initialization layer for the second stage as well. The heatmap conversion module converts the initial detected **1D** vector shape to **2D** heatmap by applying a Gaussian kernel as,

$$H = \exp\left(-\frac{(X - x_p)^2 + (Y - y_p)^2}{(2 \times \sigma^2)}\right) \tag{9}$$

where x_p and y_p are the coordinates predicted landmark and represent the center of the blob, and σ is the spread of the blob.

The concatenation of the face image and the generated **2D** heatmap from the first stage is used as the input to the next stage as in (10). These concatenated feature patches encode sufficient information about the local appearance around the current **2D** landmarks and allow the second stage to fine-tune the detected landmarks. The conversion details are illustrated in Algorithm 1.

Algorithm 1: Convert **2***D* coordinates landmarks into a heatmap

```
- Initialize w, h, $N$, $\sigma$
- Initialize $output$=$zeros[batchsize]$
- Read the landmark shape vector ($S_p$)
- Reshape $S_p$ to [$N$, 2]
for $i = 0 : batchsize$ do
    - Set $var$=$zeros(N)$
    for $j = 0 : N$ do
        - Set $X$ = [0:w]
        - Set $Y$ = [0:h]
        - Set $x_p = S_p[i,j,0]$
        - Set $y_p = S_p[i,j,0]$
        - Calculate $H$ based on (9)
        - Set $var[j]$=$H$
        - Reduce the sum of $var$ at axis 0
    end
    - Set $output[i]$=$var$
end
```

$$I_{s2} = I \oplus H. \tag{10}$$

CR-HC Loss Function

To train the CR-HC model, we used the mean absolute error (MAE) loss function L_{MAE}, which represents the sum of L_1 loss functions between the predicted landmarks and the ground truth landmarks of the model stage. L_{MAE} can be defined as

$$L_{MAE} = \sum_{1}^{s} \frac{1}{K} \sum_{i=1}^{K} \sum_{j=1}^{N} \|P_{i,j} - G_{i,j}\| \tag{11}$$

where s represents the stage number, K is the number of inputs, N is the number of landmarks, $P_{i,j}$ and $G_{i,j}$ are the detected and ground truth landmarks. Steps of the training process of the CR-HC model are provided in Algorithm 2.

Algorithm 2: Training process of the CR-HC model.

```
Require  : The training set $X$, the corresponding ground truth
           $S$, the CR-HC network $G$, and the heatmap
           coupling module.
Ensure   : Model parameters of CR-HC model
for $t = 0 : N$ do
    - Randomly, select a batch of training samples $X_t$
    - Forward $G$ by $G(X_t)$
    - Optimize $G$ by minimizing the loss function defined by
      Eq. (11)
    - Pass the predicted shape coordinates $S_p(t)$ to heat map
      coupling module and calculate the corresponding
      heatmaps $H_t$ as described in Algorithm 1
    - Input the concatenation of $X_t$ and $H_t$ to the next stage.
    - Output the $S_p(t+1)$
    - Continue until the end of epochs or the validation data
      accuracy stop of increase.
end
```

Experiments and Results

To assess the proposed method, several experiments are carried out on a variety of hard benchmarks with varying annotation schema including the Annotated Facial Landmarks in the Wild (AFLW) dataset [52], the 300 Faces in the Wild (300W) dataset [53], and the Wider Facial Landmarks in the Wild (WFLW) dataset [54]. All experiments are implemented using the Keras library on two NVIDIA Tesla K80 GPUs. Also, the training images are cropped and resized to 128 × 128 according to the provided bounding boxes and represented using RGB values. All the training dataset images are normalized by subtracting the mean image from the training set and dividing by its standard deviation. For 300W dataset image rotation, flipping and pixel shifting is applied. For AFLW and WFLW, we have used the provided training images without any data augmentation. The CR-HC model is trained from scratch using Adaptive Moment Estimation (Adam) optimization algorithm with a fixed learning rate of 0.0001 and a batch size of 32 with L_1 loss function. The number of epochs is 100, 150, and 120 for the dataset of 300W, WFLW, and AFLW, respectively. It is clear that the number of epochs is different as the challenge in each dataset is different.

Datasets

AFLW It has a large collection of images gathered from flicker, where it contains 21,997 in wild images with 25,993 faces in total. The collected images have a wide range of variety in facial appearances like pose, expression, occlusion, illumination as well as general imaging and environmental conditions. The dataset is annotated with 21 landmark coordinates. We follow the same setting used in [55] by dropping the landmarks of the ears and using only 19 landmarks. The dataset is divided into two subsets: AFLW-Full with 20,000 faces for the training phase and AFLW-Frontal with 4386 for the testing phase using the same training samples, but using only 1165 frontal faces for testing.

300W It is the most popular facial landmarks dataset, it contains five different datasets with 68 points annotation schema as LFPW, XM2VTS, AFW, IBUG, and HELEN. The same setting of [48] is applied in the current study, which is based on 3148 training images from LFPW, AFLW, and HELEN. The testing set contains all IBUg images and the test subset of HELEN and LFPW. The 135 images from IBUG are considered as the challenging test subset and 554 images from the HELEN and LFPW as the common test subset. The combination of challenge and common subsets is used as the full test set.

WFLW It is a very challenging facial landmark dataset that is introduced by [54]. It has 10,000 faces in total, 7500 for training, and 2500 for testing annotated with 98 facial points. The testing set is divided into six subsets such as occlusion, illumination, make-up, pose, expression, and blur.

Evaluation Metrics

To evaluate the proposed method and conduct a fair comparison with the state-of-the-art methods, a standard normalized mean error (NME) is considered as an evaluation metric, where

$$NME = \frac{1}{M} \sum_{i=1}^{M} \frac{\frac{1}{N} \sum_{j=1}^{N} (P_{i,j} - G_{i,j})}{d_i} \tag{12}$$

where M is the number of all tested images, and d_i is the normalization distance for 300W and WFLW. We have used an inter-ocular distance as the normalization factor, and the face size is used as the normalization factor for AFLW dataset. In addition, we used another evaluation metrics based on the failure rate at 0.1 threshold value and the area under the curve (AUC) as

$$AUC = \int_0^{th} f(e)de \tag{13}$$

where e is the normalized error, $f(e)$ denotes the cumulative error distribution function, and th denotes the upper limit of the integration for calculating the AUC.

Results

To prove the robustness of the proposed method, we conduct experiments on the three datasets using different annotated schema. Each dataset has a different number of annotated landmarks, as 19 points for AFLW, 64 points for 300W, and 98 points for WFLW. We compared the proposed method on each dataset with SDM [39], CFSS [57], ERT [37], Wing [67], LAB [54], SAN [48], TCDCN [43], 3FabRec [65], ODN [68], RCN [70], RDR [71], RCN+ [72], SHN-GCN [62], HB+SRT [60], DCNN [73], and more. The proposed CR-HC model achieves competitive results compared to these methods on the three datasets as reported in the next sections.

Performance on the AFLW Dataset

Table 2 summarizes the normalized mean error compared to the state-of-the-art methods. It is clear that the proposed method achieves a NME of 1.56% in the frontal subset, which represents about 3.70% improvement from the best previous method in [69]. The cumulative error curve (CED) is drawn in Fig. 3 for the proposed method and other methods. The proposed method achieved the highest CED curve, which differs significantly from the previous methods. The experimental results on the AFLW datasets prove that the proposed method outperforms the state-of-the-art methods by a large margin.

Performance on the 300W Dataset

To thoroughly assess the robustness of the proposed method, we conducted other experiments on the 300W three subsets (Full, Common, and Challenge). The results reported in Table 3 describe the NME of the proposed method compared to the state of the arts on the three categories. The cumulative error curve is shown in Fig. 4. It is clear that the CR-CH method achieves competitive results on the three 300W categories.

Table 2 Normalized mean error (%) on the AFLW dataset for 19 facial landmarks

Methods	Year	AFLW-Full	AFLW-Frontal
ERT [37]	2014	4.35	2.75
LBF [56]	2016	4.24	2.74
SDM [39]	2013	4.05	2.94
CFSS [57]	2015	3.92	2.69
PCPR [58]	2013	3.73	2.87
CCL [55]	2016	2.72	2.17
DAC-CSR [59]	2017	2.27	1.81
HB+SRT [60]	2021	2.26	1.64
TSR [61]	2017	2.17	-
SHN-GCN [62]	2020	2.15	-
CPM+SBR [63]	2018	2.14	-
SAN [48]	2018	1.91	1.85
DSRN [64]	2018	1.86	-
3FabRec [65]	2020	1.84	1.59
HR-LD [66]	2021	1.75	-
Wing [67]	2018	1.65	-
ODN [68]	2019	1.63	1.38
SA [69]	2019	1.62	-
Our method CR-CH	2022	**1.56**	**1.48**

Performance on the WFLW Dataset

The performance of the proposed method is also evaluated on the WFLW datasets of 98-point annotation schema. Normalized mean error, AUC at 0.1, and failure rate on the test set and six subsets are summarized in Table 4. Our approach achieves the best NME values in the test set and all subsets except the pose subset.

Table 3 Performance of the proposed method compared to other methods on the 300W test subsets for 68 facial landmarks

Methods	Year	Full	Common	Challenge	Type
PCD-CNN [74]	2018	4.44	3.67	7.62	Inter-ocular
CPM+SBR [63]	2018	4.10	3.28	7.58	Inter-ocular
RCN [70]	2016	5.41	4.67	8.44	Inter-ocular
DSRN [64]	2018	5.21	4.12	9.68	Inter-ocular
TSR [61]	2017	4.99	4.36	7.56	Inter-ocular
RCN+ [72]	2018	4.90	4.20	7.78	Inter-ocular
Two-Stage [61]	2017	4.96	4.36	7.42	Inter-ocular
Pose-Invariant [75]	2017	6.30	5.43	9.88	Inter-ocular
ODN [68]	2019	4.17	3.56	6.67	Inter-ocular
RAR [76]	2016	4.94	4.12	8.35	Inter-ocular
RDR [71]	2017	5.80	5.03	8.95	Inter-ocular
HR-LD [66]	2021	4.33	3.60	7.30	inter-ocular
Our method	**2022**	**4.20**	**3.40**	**7.48**	Inter-ocular

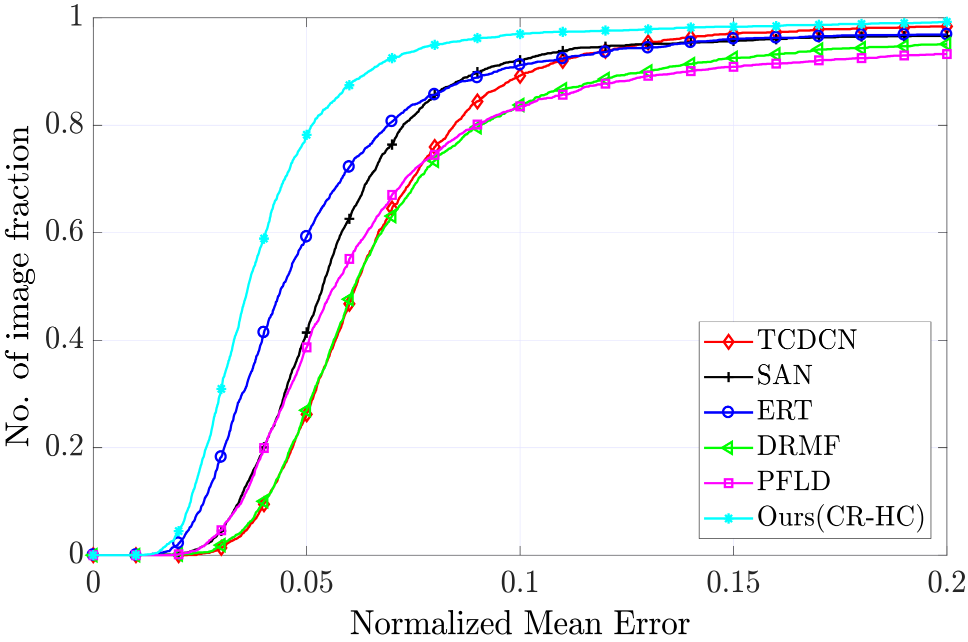

Fig. 3 Performance comparison of the cumulative error distribution curves on the AFLW dataset

Ablation Study

The proposed method consists of two main parts, the backbone convolutional neural network and the heatmap coupling module. It does not follow the same strategy of the conventional cascade coordinate regression methods. In this section, we investigate the effectiveness of the heatmap coupling module by evaluating the dataset with and without the coupling module. Figure 5 shows that the validation loss for the three datasets is decreased in the case of using the coupling module. The AFLW validation loss is decreased by 9.10% due to the use of the coupling module as shown in Fig. 5a. In the same way, the validation loss decreased for the 300W and WFLW datasets by 13.20% and 9.50% as shown in Fig. 5b, c, respectively.

The evaluated datasets have faces in uncontrolled conditions and challenge images. Figures 6, 7, and 8 show the detection results of the proposed model on the dataset of AFLW, 300W, and WFLW, respectively. The displayed images have a wide range of factors influencing the efficiency of landmarks detection, such as occlusion, head pose, illumination, and expression. The results prove the success

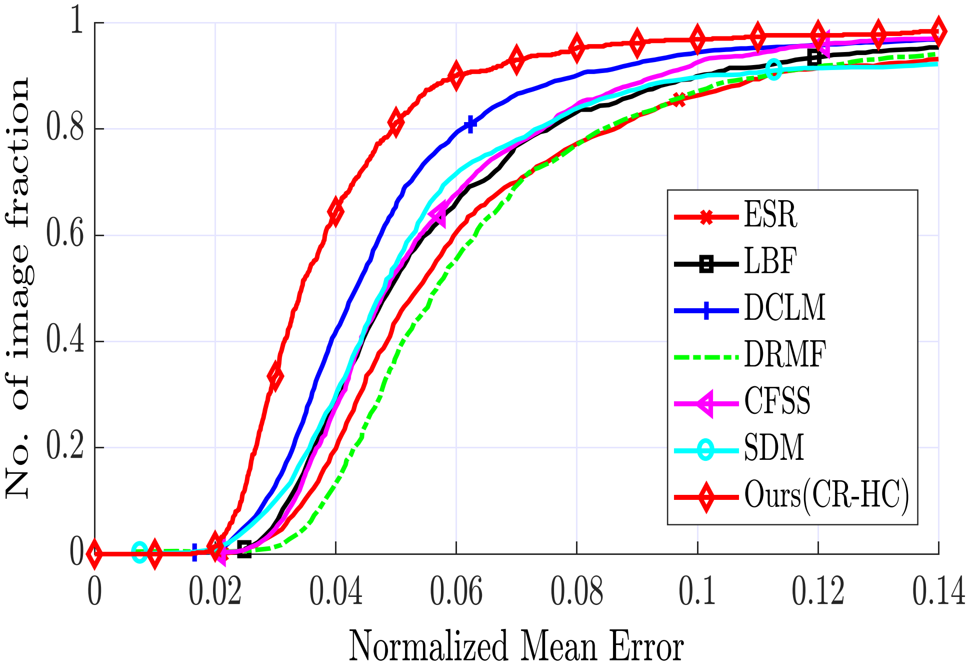

Fig. 4 Performance comparison of the cumulative error distribution curves on the 300W dataset

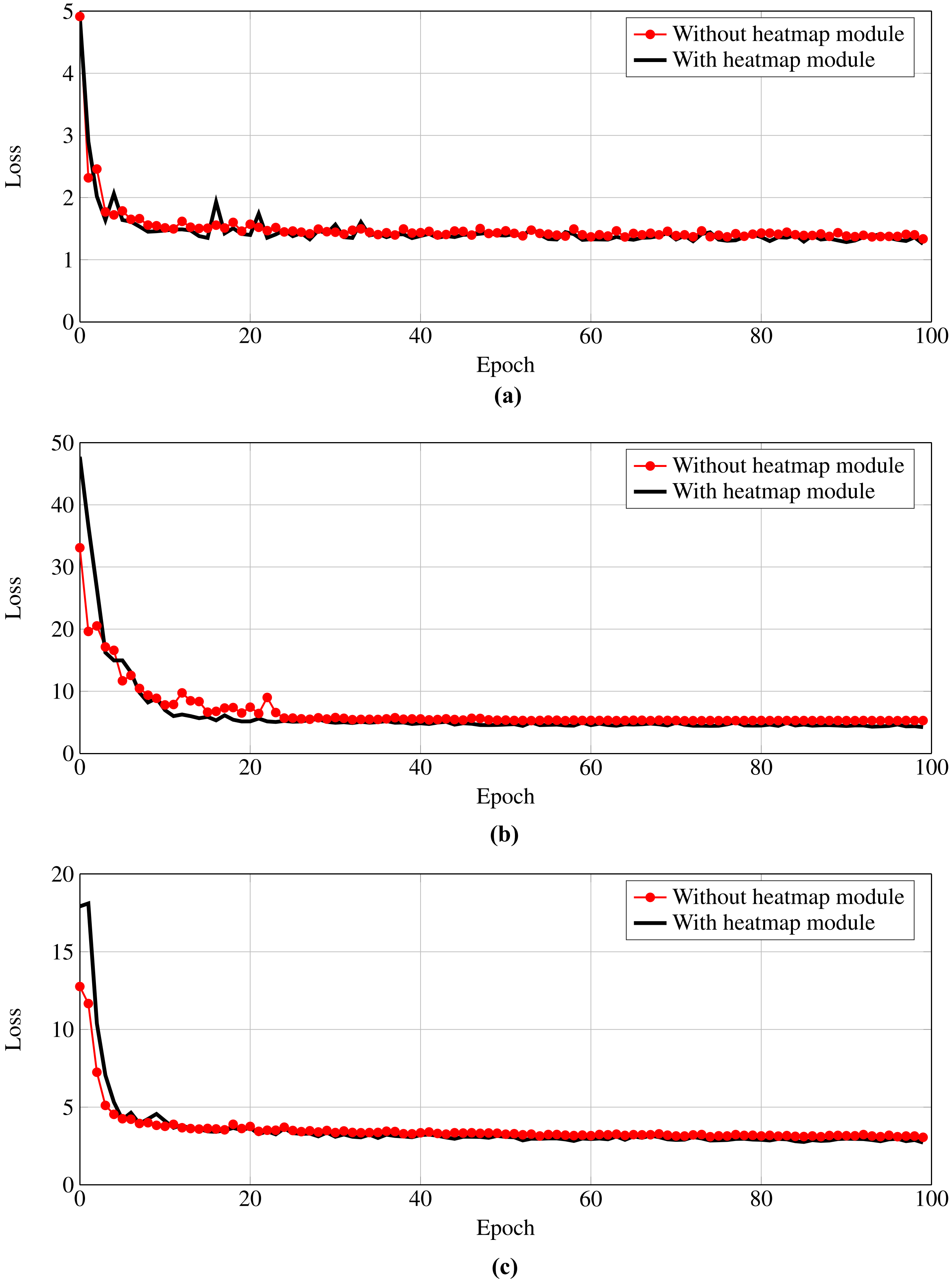

Fig. 5 The validation loss with and without the coupling module versus the epoch number for the dataset of: **a** AFLW, **b** 300W, and **c** WFLW

Table 4 Evaluation of the proposed method on the WFLW dataset compared to literature work

Metrics	Methods	All	Pose	Expr.	Illum.	Make-up	Occlusion	Blur
NME(%)(↓)	ESR [77]	11.13	25.88	11.47	10.49	11.05	13.75	12.20
	SDM [39]	10.29	24.10	11.45	9.32	9.38	13.03	11.28
	CFSS [57]	9.07	21.36	10.09	8.30	8.74	11.76	9.96
	DCNN [73]	6.08	11.54	6.78	5.73	5.98	7.33	6.88
	3FabRec [65]	5.62	10.23	6.09	5.55	5.68	6.92	6.38
	RWing [46]	5.60	9.79	6.16	5.54	6.65	7.05	6.41
	Ours	**5.53**	**11.50**	**5.80**	**5.37**	**6.05**	**6.95**	**6.28**
AUC @0.1(↑)	ESR [77]	0.2774	0.0177	0.1981	0.2953	0.2485	0.1946	0.2204
	SDM [39]	0.3002	0.0226	0.2293	0.3237	0.3125	0.2060	0.2398
	CFSS [57]	0.3659	0.0632	0.3157	0.3854	0.3691	0.2688	0.3037
	DCNN [73]	0.4551	0.1474	0.3889	0.4743	0.4494	0.3794	0.3973
	3FabRec [65]	0.4840	0.1920	0.4480	0.4960	0.4730	0.3980	0.4340
	RWing [46]	0.5182	0.2895	0.4648	0.5183	0.5102	0.4555	0.4562
	Ours	**0.5153**	**0.1990**	**0.4778**	**0.5310**	**0.4941**	**0.4378**	**0.4674**
FR@0.1(%)(↓)	ESR [77]	35.24	90.18	42.04	30.80	38.84	47.28	41.40
	SDM [39]	29.40	84.36	33.44	26.22	27.67	41.85	35.32
	CFSS [57]	20.56	66.26	23.25	17.34	21.84	32.88	23.67
	DCNN [73]	10.84	46.93	11.15	7.31	11.65	16.30	13.71
	3FabRec [65]	8.28	34.35	8.28	6.73	10.19	15.08	9.44
	RWing [46]	8.24	34.36	9.87	7.16	9.71	15.22	10.61
	Ours	**8.77**	**35.93**	**8.86**	**7.44**	**13.02**	**14.80**	**10.54**

of the proposed model to detect facial landmarks in difficult cases.

For further illustration, Fig. 9 presents a landmarks detection in the three used datasets, where row1, row2, and row3 represent the detection result in AFLW, 300W, and WFLW datasets, respectively. The results illustrate why the proposed approach might lead to inaccurate estimates in some situations. Referring to the images with indices ranging from 1 to 21, beginning in the top-left corner and going line wise, it is noticeable that when there is more than one factor affecting the distortion in the image, such as occlusion and head position as in images 1, 8, 15, and 16. The detection efficiency is affected when the color is absent from the image, leading to the overlapping of facial details, as shown in images 11, 9, and 19. Furthermore, the results are significantly impacted because only eyes are visible in images 2 and 18.

To measure the feasibility and usability of the proposed CR-HC method, the execution time is calculated and compared to other FLD methods. The execution time is calculated by computing the average execution time of 1000 images. In addition, all the compared methods are available online data source. Table 5 shows that the execution time of the proposed method is better for reliable applications compared to other FLD methods that have low execution time but have high normalized mean error on the other side.

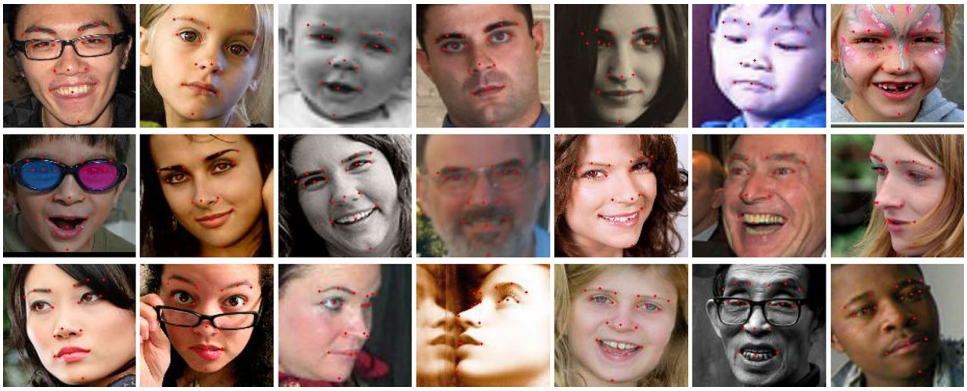

Fig. 6 Sample results of the proposed (CR-HC) method for AFLW (19 points) dataset

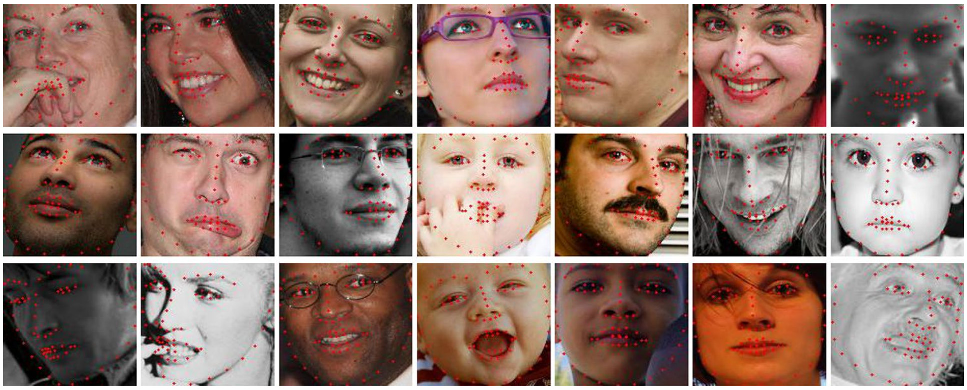

Fig. 7 Sample results of the proposed (CR-HC) model for 300W (68 points) dataset

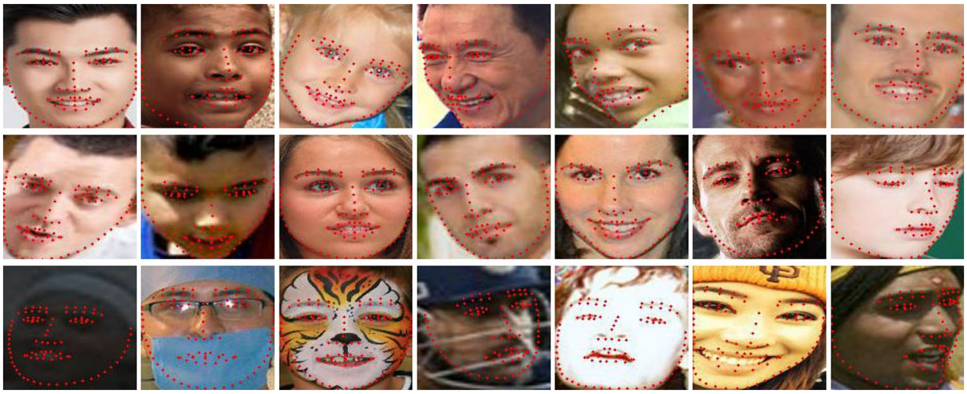

Fig. 8 Sample results of the proposed (CR-HC) method for WFLW (98 points) dataset

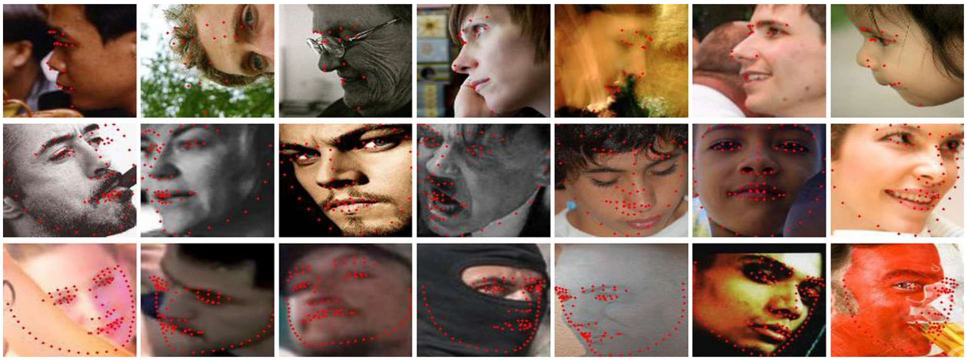

Fig. 9 Failure sample results of the proposed (CR-HC) method

Table 5 Execution time of the proposed CR-HC method compared to some other FLD methods

Method	Execution time	Device type
ERT [37]	4.18 s	CPU Intel I3-2310M
TCDCN [43]	0.12 s	CPU Intel I3-2310M
SAN [48]	26.72 s	CPU Intel I3-2310M
FAN [49]	2.68 s	Google Colab
Ours CR-HC	**3.33 s**	**Google Colab**

Conclusion

In this paper, we have presented a deep learning-based method using cascaded regression for coarse-to-fine detection of facial landmarks. The method is composed of two-stage cascaded CNNs that are coupled with a heatmap module. The first stage regresses the coordinates of landmarks of an input face image, and then it is transferred to the heatmap coupling module to convert the estimated shape to a

Gaussian heatmap. The second stage is used to refine the output by regressing the concatenation of face images and a heatmap of the estimated shape vector. The obtained results revealed that the proposed method achieved approximately 1.57% NME on the AFLW dataset, 4.30% on the 300W dataset, and 5.53% on the WFLW dataset. Thus, using the coupling heatmap module improves the detection performance distinctly. In future studies, it is possible to suggest two paths, which can increase the accuracy of FLD. First, a combination of coordinate regression as the first stage of the CR-HC model and heatmap regression network as the second stage can be done. Secondly, other large datasets can be used to train the model specifically on the WFLW dataset, which has a wide range of styling.

Data Availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study. Also, all used datasets are available free for public download.

Declarations

Ethics Approval and Consent to Participate This article does not contain any studies with human participants or animals performed by any of the authors.

Competing Interests The authors declare no competing interests.

References

1. Zou Z, Zhang X, Liu H, Li Z, Hussain A, Li J. A novel multimodal fusion network based on a joint coding model for lane line segmentation. Information Fusion. 2022;80:167–78.
2. Tanveer M, Ganaie M, Suganthan P. Ensemble of classification models with weighted functional link network. Appl Soft Comput. 2021;107:107322.
3. Fan J, Zheng P, Li S. Vision-based holistic scene understanding towards proactive human-robot collaboration. Robot Comput Integr Manuf. 2022;75:102304.
4. Huang K, Hussain A, Wang QF, Zhang R. Deep learning: fundamentals, theory and applications. vol. 2. Springer; 2019.
5. Qadeer N, Shah JH, Sharif M, Khan MA, Muhammad G, Zhang YD. Intelligent Tracking of Mechanically Thrown Objects by Industrial Catching Robot for Automated In-Plant Logistics 4.0. Sensors. 2022;22(6):2113.
6. Ma F, Gao F, Wang J, Hussain A, Zhou H. A novel biologically-inspired target detection method based on saliency analysis for synthetic aperture radar (SAR) imagery. Neurocomputing. 2020;402:66–79.
7. Cheng EJ, Chou KP, Rajora S, Jin BH, Tanveer M, Lin CT, et al. Deep sparse representation classifier for facial recognition and detection system. Pattern Recogn Lett. 2019;125:71–7.
8. Zhang Z, Xu F, Qin Z, Xie Y. Resource allocation in UAV assisted air ground intelligent inspection system. Cognitive Robotics. 2022;2:1–12.
9. Hassaballah M, Hosny KM. Recent advances in computer vision. Springer; 2019.
10. Zhao J, Xu S, Wang R, Zhang B, Guo G, Doermann D, et al. Data-adaptive binary neural networks for efficient object detection and recognition. Pattern Recognition Letters. 2022;153:239–45.
11. Zeng D, Zhao F, Shen W, Ge S. Compressing and accelerating neural network for facial point localization. Cogn Comput. 2018;10(2):359–67.
12. Zhang G, Ke Y, Zhang W, Hassaballah M. Advances and trends in video face alignment. In: Recent Advances in Computer Vision. Springer; 2019. p. 61–84.
13. Jiang C, Huang K, Zhang S, Xiao J, Niu Z, Hussain A. Towards Simple and Accurate Human Pose Estimation with Stair Network. arXiv preprint arXiv:220209115. 2022.
14. Guan Y, Fang J, Wu X. Multi-pose face recognition using Cascade Alignment Network and incremental clustering. Signal, Image and Video Processing. 2021;15(1):63–71.
15. Hassaballah M, Aly S. Face recognition: challenges, achievements and future directions. IET Computer Vision. 2015;9(4):614–26.
16. Sardar A, Umer S, Rout RK, Wang SH, Tanveer M. A Secure Face Recognition for IoT-Enabled Healthcare System. ACM Transactions on Sensor Networks (TOSN). 2022.
17. Albu F, Hagiescu D, Vladutu L, Puica MA. Neural network approaches for children's emotion recognition in intelligent learning applications. In: EDULEARN15 7th Annu Int Conf Educ New Learn Technol Barcelona, Spain, 6th-8th; 2015.
18. Qayyum A, Razzak I, Tanveer M, Mazher M. Spontaneous Facial Behavior Analysis using Deep Transformer Based Framework for Child–Computer Interaction. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM). 2022.
19. Greco A, Saggese A, Vento M, Vigilante V. Gender recognition in the wild: a robustness evaluation over corrupted images. J Ambient Intell Humaniz Comput. 2021;12(12):10461–72.
20. Qayyum A, Razzak I, Moustafa N, Mazher M. Progressive ShallowNet for large scale dynamic and spontaneous facial behaviour analysis in children. Image Vis Comput. 2022;119:104375.
21. Hu S, Shum HP, Liang X, Li FW, Aslam N. Facial reshaping operator for controllable face beautification. Expert Systems with Applications. 2021;167:114067.
22. Nan F, Jing W, Tian F, Zhang J, Chao KM, Hong Z, et al. Feature super-resolution based Facial Expression Recognition for multi-scale low-resolution images. Knowledge-Based Systems. 2022;236:107678.
23. Hannane R, Elboushaki A, Afdel K. A divide-and-conquer strategy for facial landmark detection using dual-task CNN architecture. Pattern Recognition. 2020;107:107504.
24. Hassaballah M, Murakami K, Ido S. An automatic eye detection method for gray intensity facial images. International Journal of Computer Science Issues. 2011;8(4):272.
25. Gao P, Lu K, Xue J, Shao L, Lyu J. A coarse-to-fine facial landmark detection method based on self-attention mechanism. IEEE Transactions on Multimedia. 2020;23:926–38.
26. Hassaballah M, Bekhet S, Rashed AA, Zhang G. Facial features detection and localization. In: Recent Advances in Computer Vision. Springer; 2019. p. 33–59.
27. Salem E, Hassaballah M, Mahmoud MM, Ali AMM. Facial Features Detection: A Comparative Study. In: The International Conference on Artificial Intelligence and Computer Vision. Springer; 2021. p. 402–12.
28. Jeong M, Ko BC, Kwak S, Nam JY. Driver Facial Landmark Detection in Real Driving Situations. IEEE Transactions on Circuits and Systems for Video Technology. 2018;28(10):2753–67.
29. Hassaballah M, Murakami K, Ido S. Eye and Nose Fields Detection from Gray Scale Facial Images. In: IAPR Conference on Machine Vision Applications; 2019. p. 406–9.
30. Wu Y, Ji Q. Facial landmark detection: A literature survey. Int J Comput Vis. 2019;127(2):115–42.
31. Shao X, Xing J, Lyu J, Zhou X, Shi Y, Maybank SJ. Robust face alignment via deep progressive reinitialization and adaptive error-driven learning. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022.

32. Dhamija A, Dubey R. A novel active shape model-based Deep-Neural network for age invariance face recognition. J Vis Commun Image Represent. 2022;82:103393.
33. Cootes TF, Edwards GJ, Taylor CJ. Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2001;23(6):681–5.
34. Cristinacce D, Cootes TF, et al. Feature detection and tracking with constrained local models. In: British Machine Vision Conference. vol. 1; 2006. p. 3–12.
35. Han S, Yang Z, Li Q, Chen Y. Deformed landmark fitting for sequential faces. J Vis Commun Image Represent. 2019;62:381–93.
36. Yang H, Patras I. Privileged information-based conditional regression forest for facial feature detection. In: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition. IEEE; 2018. p. 1–6.
37. Kazemi V, Sullivan J. One millisecond face alignment with an ensemble of regression trees. In: IEEE Conference on Computer Vision and Pattern Recognition; 2014. p. 1867–74.
38. Martinez B, Valstar MF. $\mathbb{L}_{2,1}$-based regression and prediction accumulation across views for robust facial landmark detection. Image Vis Comput. 2016;47:36–44.
39. Xiong X, De la Torre F. Supervised descent method and its applications to face alignment. In: IEEE Conference on Computer Vision and Pattern Recognition; 2013. p. 532–9.
40. Ren S, Cao X, Wei Y, Sun J. Face alignment at 3000 fps via regressing local binary features. In: IEEE Conference on Computer Vision and Pattern Recognition; 2014. p. 1685–92.
41. Colaco SJ, seog Han D. Deep Learning-based Facial Landmarks Localization using Compound Scaling. IEEE Access. 2022.
42. Sun Y, Wang X, Tang X. Deep convolutional network cascade for facial point detection. In: IEEE Conference on Computer Vision and Pattern Recognition; 2013. p. 3476–83.
43. Zhang Z, Luo P, Loy CC, Tang X. Facial landmark detection by deep multi-task learning. In: European Conference on Computer Vision. Springer; 2014. p. 94–108.
44. Chen Y, Yang J, Qian J. Recurrent neural network for facial landmark detection. Neurocomputing. 2017;219:26–38.
45. Zhu M, Shi D, Gao J. Branched convolutional neural networks incorporated with jacobian deep regression for facial landmark detection. Neural Networks. 2019;118:127–39.
46. Feng ZH, Kittler J, Awais M, Wu XJ. Rectified wing loss for efficient and robust facial landmark localisation with convolutional neural networks. Int J Comput Vis. 2020;128:2126–45.
47. Wan J, Lai Z, Li J, Zhou J, Gao C. Robust facial landmark detection by multiorder multiconstraint deep networks. IEEE Transactions on Neural Networks and Learning Systems. 2021.
48. Dong X, Yan Y, Ouyang W, Yang Y. Style aggregated network for facial landmark detection. In: IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 379–88.
49. Bulat A, Tzimiropoulos G. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: IEEE International Conference on Computer Vision; 2017. p. 1021–30.
50. Yin S, Wang S, Chen X, Chen E, Liang C. Attentive One-Dimensional Heatmap Regression for Facial Landmark Detection and Tracking. In: 28th ACM International Conference on Multimedia; 2020. p. 538–46.
51. Mahpod S, Das R, Maiorana E, Keller Y, Campisi P. Facial Landmarks Localization using Cascaded Neural Networks. Comput Vis Image Underst. 2021;205(1):38–59.
52. Koestinger M, Wohlhart P, Roth PM, Bischof H. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: IEEE International Conference on Computer Vision Workshops. IEEE; 2011. p. 2144–51.
53. Sagonas C, Tzimiropoulos G, Zafeiriou S, Pantic M. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: IEEE International Conference on Computer Vision Workshops; 2013. p. 397–403.
54. Wu W, Qian C, Yang S, Wang Q, Cai Y, Zhou Q. Look at boundary: A boundary-aware face alignment algorithm. In: IEEE Conference on Computer VVision and Pattern Recognition; 2018. p. 2129–38.
55. Zhu S, Li C, Loy CC, Tang X. Unconstrained face alignment via cascaded compositional learning. In: IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 3409–17.
56. Ren S, Cao X, Wei Y, Sun J. Face alignment via regressing local binary features. IEEE Transactions on Image Processing. 2016;25(3):1233–45.
57. Zhu S, Li C, Change Loy C, Tang X. Face alignment by coarse-to-fine shape searching. In: IEEE Conference on Computer Vision and Pattern Recognition; 2015. p. 4998–5006.
58. Burgos-Artizzu XP, Perona P, Dollár P. Robust face landmark estimation under occlusion. In: IEEE International Conference on Computer Vision; 2013. p. 1513–20.
59. Feng ZH, Kittler J, Christmas W, Huber P, Wu XJ. Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. In: IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 2481–90.
60. Dong X, Yang Y, Wei SE, Weng X, Sheikh Y, Yu SI. Supervision by registration and triangulation for landmark detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020.
61. Lv J, Shao X, Xing J, Cheng C, Zhou X. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In: IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 3317–26.
62. Zhang J, Hu H, Feng S. Robust facial landmark detection via heatmap-offset regression. IEEE Transactions on Image Processing. 2020;29:5050–64.
63. Dong X, Yu SI, Weng X, Wei SE, Yang Y, Sheikh Y. Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In: IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 360–8.
64. Miao X, Zhen X, Liu X, Deng C, Athitsos V, Huang H. Direct shape regression networks for end-to-end face alignment. In: IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 5040–9.
65. Browatzki B, Wallraven C. 3FabRec: Fast few-shot face alignment by reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition; 2020. p. 6110–20.
66. Kumar A, Chellappa R. Landmark detection in low resolution faces with semi-supervised learning. arXiv preprint arXiv:190713255. 2019.
67. Feng ZH, Kittler J, Awais M, Huber P, Wu XJ. Wing loss for robust facial landmark localisation with convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 2235–45.
68. Zhu M, Shi D, Zheng M, Sadiq M. Robust facial landmark detection via occlusion-adaptive deep networks. In: IEEE Conference on Computer Vision and Pattern Recognition; 2019. p. 3486–96.
69. Liu Z, Zhu X, Hu G, Guo H, Tang M, Lei Z, et al. Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection. In: IEEE Conference on Computer Vision and Pattern Recognition; 2019. p. 3467–76.
70. Honari S, Yosinski J, Vincent P, Pal C. Recombinator networks: Learning coarse-to-fine feature aggregation. In: IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 5743–52.
71. Xiao S, Feng J, Liu L, Nie X, Wang W, Yan S, et al. Recurrent 3d-2d dual learning for large-pose facial landmark detection. In: IEEE International Conference on Computer Vision; 2017. p. 1633–42.
72. Honari S, Molchanov P, Tyree S, Vincent P, Pal C, Kautz J. Improving landmark localization with semi-supervised learning. In: IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 1546–55.

73. Wu W, Wu X, Cai Y, Zhou Q. Deep coupling neural network for robust facial landmark detection. Comput Graph. 2019;82:286–94.
74. Kumar A, Chellappa R. Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment. In: IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 430–9.
75. Jourabloo A, Ye M, Liu X, Ren L. Pose-invariant face alignment with a single cnn. In: IEEE International Conference on Computer Vision; 2017. p. 3200–9.
76. Xiao S, Feng J, Xing J, Lai H, Yan S, Kassim A. Robust facial landmark detection via recurrent attentive-refinement networks. In: European Conference on Computer Vision. Springer; 2016. p. 57–72.
77. Cao X, Wei Y, Wen F, Sun J. Face alignment by explicit shape regression. Int J Comput Vis. 2014;107(2):177–90.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.