Multistage Model for Robust Face Alignment Using Deep Neural Networks

Huabin Wang¹, Rui Cheng¹, Jian Zhou¹, Liang Tao¹, and Hon Keung Kwan²
¹MOE Key Laboratory of Intelligent Computing and Signal Processing, School of Computer Science and Technology, Anhui University, Hefei, Anhui 230601 China.
²Department of Electrical and Computer Engineering, University of Windsor, Windsor, Ontario N9B 3P4 Canada.
{wanghuabin, jzhou, taoliang}@ahu.edu.cn, chengrui@ahu.edu.cn, kwanl@uwindsor.ca

Abstract—An ability to generalize unconstrained conditions such as severe occlusions and large pose variations remains a challenging goal to achieve in face alignment. In this paper, a multistage model based on deep neural networks is proposed which takes advantage of spatial transformer networks, hourglass networks and exemplar-based shape constraints. First, a spatial transformer - generative adversarial network which consists of convolutional layers and residual units is utilized to solve the initialization issues caused by face detectors, such as rotation and scale variations, to obtain improved face bounding boxes for face alignment. Then, stacked hourglass network is employed to obtain preliminary locations of landmarks as well as their corresponding scores. In addition, an exemplar-based shape dictionary is designed to determine landmarks with low scores based on those with high scores. By incorporating face shape constraints, misaligned landmarks caused by occlusions or cluttered backgrounds can be considerably improved. Extensive experiments based on benchmark datasets are performed to demonstrate the superior performance of the proposed method over other state-of-the-art methods.

Index Terms—Face alignment, facial landmark detection, multistage model, spatial transformer generative adversarial networks, generative deep neural networks, discriminative deep neural networks, stacked hourglass networks, convolutional neural networks, residual units, deep residual networks, exemplar-based shape constraints, K-means algorithm.

I. INTRODUCTION

FACE alignment (or facial landmark detection) aims to locate a set of predefined human facial landmarks, such as the corners of the eyes, the eyebrows, and the tip of the nose for high-level vision tasks, such as face recognition [1], expression recognition [2], facial animation [3] and 3D face modelling [4]. Although considerable progress has been made, face alignment is still challenging due to large-view face variations, lighting conditions, complex expressions and partial occlusions.

Recently, progresses have been made by convolutional neural networks (CNNs) in semantic segmentation [5] and in human pose estimation and face alignment based on heatmap regression [6]. The hourglass network [6] offers a method for human pose estimation. The model utilizes repeated down-sampled and up-sampled modules to extract features across multiple scales. The hourglass network has been introduced to face alignment task and achieved efficient performance. However, existing methods are still inefficient in modelling face structural priors, the performance of these methods degrades severely when face images suffer from heavy occlusion, and this problem is challenging to address since occlusion is common and diverse in reality.

Several typical face alignment models have attempted to address faces under partial occlusions. Robust cascaded pose regression (RCPR) [7] is the first method that simultaneously detects landmarks and estimates occlusions. In this method, the face is divided into a 3 × 3 grid for each regression stage, and only one non-occluded face region is used to predict the location of the landmarks. The work in [8] proposed a unified framework that combines landmark localization and visibility estimation, which focuses more on landmarks with high visibility probabilities and iteratively updates landmark locations and landmark visibility probabilities. Xing et al. [9] considered the regression procedure as a sparse coding problem by learning two dictionaries: one is the face appearance dictionary, the other is the face shape dictionary. With two relational dictionaries, the occluded face appearance is restored, and the influence of the occluded landmarks is suppressed. Liu et al. [10] utilized shape-indexed appearance to estimate the occlusion level of each landmark, and the face shape is reconstructed by similar shapes from the exemplar-based shape dictionary. Although these methods have shown superior performance in detecting occluded landmarks, they still suffer from poor scalability and robustness. The first limitation is the lack of large-scale ground truth occlusion annotation for natural images. The task of providing occlusion annotation is often time consuming, involving a considerable amount of tedious manual work. Additionally, due to the inherent complex variations in human facial appearance in unconstrained environments, it is difficult to recover the occluded appearance using face appearance dictionary.

Another challenge is the initialization issue of face images derived from face detectors, which has drawn little attention in previous studies. The pre-processing step of face alignment is to crop face rectangles through a face detector. However, due to severe occlusion or blur, the face detector may not produce an appropriate face rectangle. As Ren et al. noted in [11], if the initial images have different scale and rotation variations, the performance of many face alignment methods would be severely degraded. It will be useful if an algorithm could produce canonical face poses with the same scales and centre
This work is built on top of [14] with four major contributions as listed at the end of Section I.

In summary, we make the following contributions to the face alignment task:

1) A spatial transformer - generative adversarial network is proposed to produce promising initial face images for face alignment.

2) Based on the intensity of the heatmaps obtained by a two-stage hourglass network, a scoring scheme is designed to measure the quality of predicted landmarks locations, which can estimate the occlusion level of each landmark and distinguish the aligned landmarks from misaligned landmarks.

3) An exemplar-based shape dictionary is employed to impose geometric constraints. The landmarks with high scores are used to search similar shapes from dictionary, and the landmarks with low scores are refined by shape reconstruction using similar shapes.

4) Experiment results on several benchmark datasets (300-W, COFW and WFLW) show that the proposed multistage model outperforms most recent face alignment methods, especially for faces with difficult scenarios such as large pose, lighting and occlusion, etc.

II. RELATED WORK

In this section, we first review the development of face alignment, and then briefly review STNs.

A. Face Alignment

Face alignment methods can be generally classified into three categories: discriminative fitting, cascaded shape regression, and deep learning.

Since facial shape and facial appearance are deformable structured objects, methods based on discriminative fitting
typically model facial structures by learning shape and appearance variation models. According to the difference in facial representations, these methods can be divided into two categories: one is the holistic-based representation, such as active appearance model (AAM) [17], the other is part-based representation, such as active shape model (ASM) [18], constrained local model (CLM) [19], Gauss-Newton deformable part model (GN-DPM) [20]. These methods typically require an iterative process to find the optimal parameter configuration for a given face, thus it is time-consuming and prone to fall into local minima. Moreover, due to the limited capacity of parametric models, such methods are sensitive to occlusion and large pose variation.

Methods based on cascaded shape regression were popular in face alignment before the advent of deep learning. These approaches are based on a multistage framework, and each stage refines the position of predicted landmarks in a coarse-to-fine manner. Specifically, a weak regressor is utilized in each stage to model the relation between the image feature and the shape increment. Cootes et al. [21] proposed an efficient method that combines random forest regression and a statistical shape model. The supervised descent method (SDM) [22] focuses on solving the optimization problem of the least squares method. Ren et al. [11] proposed learning local binary features around local patches using random forest regression, which was faster than existing methods. In [23], a projective invariant is designed for modelling the intrinsic structure of human faces and combined it with cascade regression methods. The regression-based approach mentioned above employs the handcrafted feature descriptors (e.g., SIFT [22], HoG [24], or random forest/fern descriptors [11]) to extract facial texture information. It is clear that conventional cascaded regression methods have yielded drastic improvements based on standard benchmarks such as 300-W [25]. However, most of these methods are sensitive to initialized shapes, due to the limitations of handcrafted features.

Recently, CNNs have made a series of breakthroughs in many visual analysis tasks such as image classification [26], semantic segmentation [5], and human pose estimation [6]. The application of CNNs greatly boosts the performance of face alignment. CNN-based methods can be generally classified into two categories: coordinate regression methods [27]–[29] and heatmap regression methods [30]–[35]. The difference between the two categories is that the former directly regresses landmark coordinates with a network, and the latter first learns a mapping function from image to likelihood heatmaps, and chooses the location with the highest response value in the heatmap as the predicted location. Sun et al. [27] first introduced CNNs to the face alignment field, and cascaded three CNNs to detect facial landmarks in a multistage manner. The method in [28] jointly learns landmark localization and correlated recognition tasks, such as facial attributes and expressions. Xiao et al. [29] proposed a framework that leverages the advantages of CNNs and recurrent neural networks (RNNs). The feature extraction stage is replaced with a CNN, and the fitting stage is replaced with an RNN. Weng et al. [36] proposed an exemplar-based cascaded auto-encoder network for real-time face alignment.

These coordinate regression methods can directly detect the coordinates of landmarks and do not require post-processing operations. However, since coordinate regression methods are predicted landmarks from dense layers that contain high-level semantic information but lack the details of the facial texture, result in limitations in real-world scenarios, such as occlusion, large poses, and other uncontrolled conditions. Kowalski et al. [30] first introduced the idea of heatmaps to cascaded CNNs. They generated heatmaps based on the predicted coordinates of the previous stage, and then combined the original image as an input for the next stage. In [32], a binary hourglass network with a multi-scale feature fusion residual module is developed to boost performance for 2D and 3D face alignment. Deng et al. [33] employed affine transformation to remove rotation and scale variations in facial images and then detected landmarks through hourglass networks. In [34], the concept of boundary heatmap is introduced as a facial geometry. Valle et al. [35] combined a CNN and ensemble of regression trees (ERT) to enhance computational efficiency. Although heatmap regression methods represented by hourglass networks show excellent performance, there are still many limitations for hourglass networks to model the geometric structure of the human face.

B. Spatial Transformer Network

CNNs achieve excellent performance in local feature representation. However, CNNs still lack the ability to be spatially invariant to the input image. Jaderberg et al. [15] first presented STN that explicitly learns invariance to translation, scale and rotation. Benefiting from STN, they achieved state-of-the-art performance in several image classification tasks, such as MNIST [37] digit classification. STN allows a neural network to learn how to perform spatial transformations on an input image to enhance the geometric invariance of the model. In [38], an STN was embedded in cascaded CNNs, to jointly learn spatial transformation and landmark localization for face detection. Similarly, the work of [12] embedded an STN as a subnet to obtain an improved initial image for face landmark localization. In [39], STN is applied to the task of image composition, and an STN is embedded in the generator of the generative adversarial network (GAN) for warping a specific object of a given image and placing it in the scene image. Apparently original STN is robust to handling the spatial transformation of simple objects, such as handwritten digits. Due to the complex variations of faces in uncontrolled conditions, the original STN has difficulty in robustly providing accurate spatial transformations.

III. METHOD

As illustrated in Fig. 1, MSM consists of three pivotal steps: GAN-based spatial transformation, CNN-based landmark detection and exemplar-based shape reconstruction. In this section, MSM is described in detail.

A. Spatial Transformer - Generative Adversarial Network

Recent studies [11], [12] have shown that the pre-processing of face images is critical to face alignment tasks. If the
initialized image has a large pose or excessive unnecessary background, the accuracy of landmark localization is greatly reduced. There are two typical methods for facial image pre-processing: one is based on affine transformation, and the other is based on STNs. Affine transformation methods first detect several fiducial key points and then calculate the parameters of affine transformation by Procrustes analysis based on located key points and the key points of the mean face shape. It is obvious that affine transformation methods have the same limitations as the conventional face alignment algorithm, regarding sensitivity to occlusion and blur. STN-based methods explicitly learn image warping without key point detection, which is more flexible and robust than the affine transformation approach. Nonetheless, due to the complexity of the human face in nature, it is challenging to regress accurate transformation parameters using the basic STN model.

To improve the robustness of STN [15] to handling complex face images, adversarial learning is introduced. As shown in Fig. 2, the proposed spatial transformer - generative adversarial network (ST-GAN) consists of two parts: a generative deep neural network (GDNN) and a discriminative deep neural network (DDNN). Similar to original STN [15], the generative deep neural network consists of three main components: a localization network, a grid generator and a sampler. The localization network is realized by a convolutional network consisting of 11 convolutional layers with different strides. The overall configuration of the proposed GDNN and DDNN are listed in Table I and Table II, respectively. The size of input of GDNN is 128 × 128. Each of the first 9 convolutional layers of the GDNN is of size 3 × 3 with different strides. At the end, a 4 × 4 global average pooling layer and a 1 × 1 convolutional layer are utilized to regress the transformation matrix $\theta$. For 2D affine transformation, the transformation matrix $\theta$ is selected to be a 2 by 3 matrix.

$$\theta = \begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{pmatrix}$$ (1)

The grid generator generates a grid $R = \{g_i\}, g_i = [x_i, y_i]$ in the input image corresponding to each pixel $i$ from the output image. The sampler uses the transformation matrix $\theta$ and applies it to the input image. Specifically, assuming $(x^*_i, y^*_i)$ are the source coordinates of the $i$-th of the input image and $(x^g_i, y^g_i)$ are the target coordinates of the $i$-th of the output image, the transformation procedure is defined as follows.

$$\begin{pmatrix} x^g_i \\ y^g_i \end{pmatrix} = \theta(g) = \begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{pmatrix} \begin{pmatrix} x^*_i \\ y^*_i \\ 1 \end{pmatrix}$$ (2)

Similar to [12], supervised learning is applied to train affine transformation parameters. As shown in Table II, the size of the input of DDNN is 128 × 128, and the output is a scalar representing the possibilities. Each of the first 6 convolutional layers is of size 4 × 4 with stride 2, the convolutional layer 7 is of size 2 × 2 with stride 1. The loss function of discriminator DDNN is defined as follows (for simplicity, GDNN is denoted as $G$, DDNN is denoted as $D$):

$$L_D = E[\log D(I_{real})] + E[\log (1 - D(G(I_{fake})))$$ (3)

where $I_{real}$ refers to real sample which is the ground truth image without rotation, scale and unnecessary background. $I_{fake}$ refers to noise sample which is a designed facial image with rotation, scale and unnecessary background. $E$ represents the expectation. The discriminator learns to predict the ground truth facial image as one while predicting the generated facial image as zero. With DDNN, the adversarial loss can be defined as follows:

$$L_A = E[\log(1 - D(G(I_{fake})))$$ (4)

![Fig. 2: Architecture of spatial transformer - generative adversarial network (ST-GAN). The generative deep neural network (GDNN) is used to generate the transformation matrix $\theta$. The discriminative deep neural network (DDNN) is used to determine whether the generated face image is “real”, which means a canonical face without unnecessary background.](image-url)
designed network with the capability of capturing information with a repeated encoder and decoder architecture, has proven to be effective for detecting visible landmarks. Stacked hourglass network [6], which is a residual neural network, is an effective method for detecting landmarks to converge. Thus, the objective of the preliminary landmark detection is to predict its output is fake. The whole training process is organized by a minmax game in which $D$ tries to maximize the probability it correctly classifies the face pose is canonical or not (i.e. real or fake), and $G$ tries to minimize the probability that $D$ will predict its output is fake. The whole training process is summarized in Algorithm 1.

The loss function of generator $G$ is defined as
\[ \mathcal{L}_G = a||\hat{\theta} - \theta^*|| + b\mathcal{L}_A \] (5)
where $\hat{\theta}$ is the parameter regressed by GDNN and $\theta^*$ is the ground truth transformation parameter. The hyper-parameters $a$ and $b$ are used to balance different losses. Thus, GDNN is optimized to fool discriminator DDNN by regressing more accurate parameter that will improve the learning of the spatial transformation. The final objective function can be expressed as follows.
\[ \arg\min_G \max_D (\mathcal{L}_G + \mathcal{L}_D) \] (6)
In this way, the generator $G$ and the discriminator $D$ play a minimax game in which $D$ tries to maximize the probability it correctly classifies the face pose is canonical or not (i.e. real or fake), and $G$ tries to minimize the probability that $D$ will predict its output is fake. The whole training process is summarized in Algorithm 1.

Algorithm 1 Training process of ST-GAN.

Require: Training images $I_{fake}$, the corresponding ground-truth image $I_{real}$, the generator $G$, the discriminator $D$.
1: Forward $G$ by $G(I)$, and optimize $G$ according to Eq. 5;
2: Forward $D$ by $D(I_{real})$ and optimize $D$ by maximizing the first term of $\mathcal{L}_D$ defined in Eq. 3;
3: Forward $D$ by $D(I_{fake})$ and optimize $D$ by maximizing the second term of $\mathcal{L}_D$ defined in Eq. 3;
4: Optimize $G$ by Eq. 6;
5: Go back to Step 1 until the accuracy of the validation set stop increasing;
6: return $G$.

The loss function for intermediate supervision at the end of each hourglass module, 4) It can produce pixel-wise predictions of the same resolution as the input image. Recently, many work adopted four or eight hourglass modules as network backbone, but such strategy are computationally expensive for real-time applications.

To achieve a good trade-off between performance and efficiency, a network based on two hourglass modules is designed. Residual unit [26] are used as the building blocks in the hourglass network. Fig. 4 gives the detail of a 3-layer residual unit. A residual block can be expressed as follows:
\[ x_{n+1} = x_n + F(x_n, W_n) \] (7)
Where $x_{n+1}$ and $x_n$ are the output and input feature maps of the $n$-th block, $W_n$ denotes the weights of convolutional layers. $F$ consists of batch normalization, ReLU is used for non linearity function, two $1 \times 1$ convolutional layers and a $3 \times 3$ convolutional layer, with an $1 \times 1$ skip convolutional layer are used to match different channels of input and output feature maps. Stacked residual units can increase feature channels and extract high-level discriminative features. First, we give an overview of the network architecture. As shown in Fig. 3, the input of the network is a face image normalized by the previous ST-GAN with a spatial resolution of $128 \times 128$, followed by two $3 \times 3$ convolutional layers to increase the number of feature channels and a max pooling layer to decrease the resolution from $128$ to $64$, through a $3 \times 3$ convolutional layer and a residual unit, the number of channels is increased to $256$, then the feature maps with $256$ channels and $64 \times 64$ resolution are fed to the hourglass module. The hourglass module consists of a four-layer recursive structure, and each level consists of a downsampling layer, residual units, a skip connection layer and a deconvolutional layer. Considering computational costs, $64 \times 64$ resolution is used in the hourglass module. Unlike the original hourglass module [6]
which uses upsampling layer to recover the size of the feature maps, deconvolution [40] is introduced to replace upsampling layers to better maintain spatial semantic information. Batch normalization is performed before all convolutional layers to accelerate convergence except for the first convolutional layer with $3 \times 3$ kernels. ReLU is used as an activation function.

For an image $I$, this network is trained to obtain $L$ heatmaps $H(I)$, where $L$ is the total number of landmarks for each face. The location of each predicted landmark is decoded from corresponding heatmap by taking the location with the maximum value as follows:

$$c(l) = \arg \max H^l(I)$$

(8)

where $l$ is the index of the landmark and the corresponding heatmap, $c(l)$ gives the coordinate of the $l$-th landmark. Some examples output by this network are shown in Fig. 5. Note that visible landmarks can be precisely located; however, these results may not have a biological human facial shape since occluded landmarks were not detected. In addition, the response heatmaps of visible landmarks are more focused than those of occluded landmarks. It is challenging to decode corrected positions from scattered heatmaps, which is a limitation of the heatmap regression-based method.

To review the definition of a heatmap. During training, a ground truth heatmap for one landmark is created by putting a Gaussian peak at the ground truth location of a landmark, and the intensity decreases with the distance to the closest landmark. Motivated by a recent study [10] that used shape-indexed appearance to estimate the occlusion level of each landmark, the intensity of the heatmap is employed to estimate location quality and further distinguish reliable landmarks and missing landmarks. In detail, each landmark is weighted based on the corresponding intensity values in the heatmaps. Thus, more reliable landmarks with strong local information are assigned high weights. The landmarks under occlusion are assigned low weights. The process of assigning weight can be expressed by the following equation:

$$w_l = \frac{\sum_{k=X_l-r}^{X_l+r} \sum_{t=-r}^{t+r} score_v(k,t)}{(2 \times r + 1)^2}$$

(9)

where $score_v(k,t)$ is the value of coordinate $(k,t)$ in the $l$-th heatmap, $r$ determines the size of the rectangle used to calculate the score. The coordinate $(X_l, Y_l)$ gives the predicted location of the $l$-th landmark. Based on the assigned weight, the predicted landmarks can be classified into two categories: reliable landmarks and misaligned landmarks. The coordinate and weights of reliable landmarks act as initial information for the following shape refinement stage.

C. Exemplar-based Shape Reconstruction

Deep convolutional neural networks have a strong capacity for local feature representation, thus the visible landmarks can be effectively located through the first two stages. However, a large number of parameters can easily lead to network overfitting, especially for limited training samples. In addition, CNNs still lack the ability to model the geometric structure of the human face, resulting in sensitivity to occlusion. In contrast, human vision is capable of predicting face shapes by utilizing geometric constraints. Motivated by this ability, these misaligned landmarks can be refined by similar face shapes in the training samples, and this approach is feasible and simple. To this end, following [10], [41], sparse shape constraints are incorporated to correct the misaligned landmarks. The sparse shape model is a popular method of imposing shape priors, it can refine the gross error and maintains shape detail at the same time. This feature allows the model to be perfectly integrated with CNNs. The objective of the sparse shape model can be formulated as follows:

$$\arg \min \| S - D_s \alpha \|_2 + \lambda \| \alpha \|_2$$

(10)

where $S$ is a $2L \times 1$ vector with $L$ landmark coordinates of the predicted normalized shape. $D_s$ is an $N \times 2L$ matrix, that is a shape dictionary with a sample size of $N$. $\alpha$ is the shape reconstruction coefficient, and $\lambda$ is the regularization parameter. As Liu et al. noted in [10], the traditional sparse shape model treats all landmarks equally, causing the error from corrupted landmarks spread to other aligned landmarks, and harms the convergence of the model. In other words, incorrect reconstruction targets lead the sparse shape constraint to produce incorrect shapes. Different from [10], only the accurately aligned landmarks which were assigned high weights are used to search for similar shapes from a dictionary. As shown in Fig. 6, this part of the facial shape, which consists of only reliable landmarks, is our reconstruction target.

After the first two stages, the preliminary coordinates and weight of each landmark can be determined. Then a threshold $T$ is set to distinguish reliable landmarks and misaligned landmarks, Thus, for each shape $S$ we obtained a binary vector $V$. If the $l$-th component of $V$ is 1, then the $l$-th landmark is considered reliable. Based on reliable landmarks, the search process can be formulated as follows:

$$\min \| V^* S - (V^* S \odot V^* D_S) \alpha \|_2^2$$

(11)

where $V^* = diag(V)$. The goal of $V^*$ is to force the search process to neglect misaligned landmarks and emphasize landmarks with high weights. $\odot$ indicates searching for the most similar shape in the dictionary. $(V^* S \odot V^* D_S)$ is used to search for the $k$ nearest exemplar shapes of $V^* S$ from the adaptive shape dictionary $V^* D_s$. Then the misaligned part shape can be reconstructed by the $k$ nearest shapes.
and the reconstruction coefficients can be simply computed by the least squares method. However, searching all training samples is time consuming, especially for a large training set. Furthermore, there are many similar face shapes that are redundant. Thus, K-means algorithm is applied to all training shapes to obtain \( N \) representative face shapes, which form a compact shape dictionary \( D_S \). Searching from \( D_S \) will be more effective. The shape reconstruction procedure is shown in Fig. 6. The whole process of the proposed multistage model is summarized in Algorithm 2.

**Algorithm 2 Multistage Model**

**Require:** Face image \( I \), face rectangle \( R \), shape dictionary \( D_S \), threshold \( T \).

1. Crop \( I \) according to \( R \), get facial part image \( I_R \).
2. Feed \( I_R \) to ST-GAN, get normalize face image \( I_N \) and transform parameter \( \theta \).
3. Feed \( I_N \) to stacked hourglass network, get preliminary face shape \( S \).
4. Calculate weight \( w_i \) by Eq. 9 for each landmark.
5. for \( i = 1 \) to \( L \) do
6. \( w_i = w_i \times \max_{i=1,2,...,L}(w_i) \)
7. if \( w_i > T \) then \( V_i = 1 \)
8. else \( V_i = 0 \)
9. end if
10. end
11. \( V^* = \text{diag}(V) \).
12. Shape reconstruction via
\[
\text{arg min}_\alpha \| V^* S - (V^* S \odot V^* D_S) \alpha \|^2_2.
\]
13. Final shape \( S_F = D_S \alpha \).
14. Affine to original resolution by \( S_G = \theta^{-1} S_F \).

**IV. EXPERIMENTS**

In this section, we conduct extensive experiments and analysis to show the effectiveness of the proposed method. The following paragraphs describe the datasets, implementation details, experimental results and ablation study.

**A. Datasets**

Our method is evaluated on several challenging datasets including 300-W, COFW and WFLW.

1) **300-W** [25]: 300-W is currently the most widely used dataset. It was created from four datasets including the AFW [42], LFPW [43], HELEN [44] and IBUG [25] dataset, each face image is annotated with 68 landmarks. The training set consists of the AFW, LFPW training set and HELEN training set, resulting in a total of 3148 images. The test set consists of three parts: the common set, challenge set and full set. The common set consists of the LPFW test set and HELEN test set, resulting in a total of 554 images. The challenge set, which is the IBUG dataset, contains 135 images. The full set consists of a common set and challenge set containing 689 images.

2) **300-W private test set** [45]: The 300-W private test set was introduced after the 300-W dataset and was used for the 300-W Challenge benchmark. It consists of 300 indoor images and 300 outdoor images, each image was annotated 68 landmarks using the same annotation scheme as the one of 300-W.

3) **COFW** [7]: The COFW dataset focuses on occlusion in nature. The training set consists of 1345 images, the testing set consists of 507 faces with a wide range of occlusion patterns, and each face is annotated with 29 landmarks. In our experiment we use reannotated version [46] of the 68 landmarks for comparison to other approaches.

4) **WFLW** [34]: WFLW is considered the most challenging dataset. It contains 10000 faces (7500 for training and 2500 for testing) with 98 fully manually annotated landmarks and corresponding facial bounding boxes. Compared to the above datasets, WFLW includes rich attribute annotations, such as occlusion, pose, make-up, blur and illumination attribute information.

**B. Evaluation Metrics**

Similar to previous methods, we use the normalized root mean squared error (NRMSE), cumulative errors distribution (CED) curve, area under the curve (AUC) and failure rate to measure the landmark location error.

\[
NRMSE = \frac{1}{N} \sum_i \frac{1}{L} \sum_j |P_{ij} - G_{ij}| \quad (12)
\]

where \( N \) is the number of total images, \( L \) is the number of total landmarks for a given face, and \( P_{ij} \) and \( G_{ij} \) denote the predicted and ground truth locations, respectively. \( d_i \) is
the normalization parameter. The experiment results using different definitions of \( d_i \): the distance between the eye centres (inter-ocular) and the distance between the outer eye corners (inter-ocular).

For the 300-W, 300-W test set and COFW dataset, image with an NRMSE (inter-ocular) of 0.08 or greater is considered a failure. For the WFLW dataset, following [34], image with an NRMSE (inter-ocular) of 0.1 or greater is considered a failure.

C. Implementation Details

We independently trained three models: ST-GAN, stacked hourglass network and face shape dictionary. For ST-GAN, the faces are cropped by the provided bounding boxes and resized to 128×128 resolution. Data augmentation is applied by random flipping, rotation (between ±30°), scaling (between ±10%) and colour jittering. The network is optimized by Adam stochastic optimization with an initial learning rate of 0.0005 and reduced by half after 400 epochs. In total, 1000 epochs are used in training. The mini batch size is set to 16. The stacked hourglass network was trained following a similar procedure, and the difference is that the input images of the network are cropped by ground truth bounding boxes, training is applied for a total of 300 epochs. The learning rate is reduced to half after 100 epochs. Both networks were implemented in PyTorch.

Fig. 7: Face shape reconstruction based on \( k \) nearest exemplar shapes in dictionary with size \( N \). The results are obtained using COFW dataset.

In the face shape dictionary training procedure, the 300-W training set and semifrontal face of the Menpo [47] dataset are used to train 68-point face shape dictionaries. Additionally, the WFLW training set is used to train 98-point face shape dictionaries. First, affine transformation is performed with the ground truth coordinates of the pupil and the coordinates of the midpoint to make the face canonical. Then, the face shapes are normalized by converting the coordinates of each landmark to a 128×128 space. K-means algorithm is utilized to cluster normalized face shapes to reduce spatial redundancy and improve the computational efficiency. As shown in Fig. 7, we tested different dictionary sizes \( N \) and different numbers \( k \) of face shapes for reconstruction. Finally \( N \) and \( k \) are set as 500 and 100, respectively. Therefore, the face shapes are reconstructed by 100 most similar shapes in dictionary with size 500. The reconstruction coefficients are computed by the least squares method and ridge regression. The regularization parameter of ridge regression is set to 60. In Eq. 5, \( a \) and \( b \) are set to 1 and 0.5, respectively.

| Method     | Year | Common Subset | Challenge Subset | Fullset |
|------------|------|---------------|-----------------|--------|
| NRMSE (inter-ocular) (%) |
| LBF [11]   | 2014 | 4.95          | 11.98           | 6.32   |
| TCDCN [48] | 2014 | 4.80          | 8.60            | 5.54   |
| CFSS [49]  | 2015 | 4.73          | 9.98            | 5.76   |
| MDM [50]   | 2016 | 4.83          | 10.14           | 5.88   |
| RAR [29]   | 2016 | 4.12          | 7.00            | 4.68   |
| DAN [30]   | 2017 | 4.42          | 7.57            | 5.03   |
| TSR [12]   | 2017 | 4.36          | 7.56            | 4.99   |
| SHN [13]   | 2017 | 4.12          | 7.00            | 4.68   |
| LAB [34]   | 2018 | 3.42          | 6.98            | 4.12   |
| DCFE [35]  | 2018 | 3.83          | 7.54            | 4.55   |
| 3DDE [51]  | 2019 | 3.73          | 7.10            | 4.39   |
| AGCFN [10] | 2019 | 3.73          | 7.24            | 4.42   |
| MSM        | 2019 | 3.74          | 6.97            | 4.38   |

Our model is implemented on Ubuntu 18.04 with a NVIDIA GTX1080 (8GB) GPU and an Intel Core 7500 CPU @3.4 GHz×4. Training the ST-GAN and stacked hourglass network took around 8 hours and 6 hours respectively. The Python implementation process images at 14 FPS on average, the CNN part (the ST-GAN and stacked hourglass network) took around 50 ms and the shape reconstruction took around 20 ms per image.

D. Experiment using 300-W dataset

Many existing methods have established a series of impressive results on this dataset. In Table III, we compare our results with LBF [11], TCDCN [48], CFSS [49], MDM [50], RAR [29], DAN [30], TSR [12], SHN [13], LAB [34], DCFE [35], 3DDE [51], PCD-CNN [52], SAN [53], DeCaFA [55], AGCFN [56] and ODN [54] are also used in Table III.

First, we report the NRMSE results on 300-W dataset of the proposed MSM method and those of other methods in Table III. For the Challenge Subset of 300-W, the MSM achieves an inter-pupils NRMSE of 6.97% and an inter-ocular NRMSE of 4.83%. This demonstrates the MSM is robust to handling face under difficult scenarios such as large pose, lighting and occlusion, etc. For the Common Subset and Fullset of 300-W, the inter-pupils NRMSE values of LAB is slightly better than those of the MSM. However, the LAB is much more computational expensive due to a network architecture using eight stacked hourglass modules versus two stacked hourglass modules in the MSM. For the Common Subset and Fullset of 300-W, comparable inter-ocular NRMSE values are obtained by the 3DDE using a UNet-based network and MSM using two stacked hourglass modules in which MSM obtained slightly higher and slightly lower NRMSE values respectively in the
Fig. 8: MSM example outputs using 300-W dataset. For clarity of illustration, detected key points are connected to show dotted face shapes.

| Method   | NRMSE (%) | Failure (%) | AUC   |
|----------|-----------|-------------|-------|
| CFSS     | -         | 12.30       | 0.4132|
| MDM      | 5.05      | 6.80        | 0.4532|
| DAN      | 4.30      | 2.67        | 0.4700|
| SHN      | 4.05      | -           | -     |
| DCFE     | 3.88      | 1.83        | 0.5242|
| AGCFN    | 3.82      | 1.60        | 0.5252|
| MSM      | 3.81      | 1.50        | 0.5262|

Table IV: Inter-ocular NRMSE (%), failure rate (%) and AUC of face alignment results using 300-W private test set.

The MSM results using 300-W dataset. For the 300-W private test set, the comparison of NRMSE, failure rate and AUC in Table IV indicate that the MSM outperforms all other methods in NRMSE values, failure rate and AUC with the exception that the DCFE achieved an AUC of 0.5242 versus the MSM of 0.5262.

Fig. 9: CED curves of face alignment results using 300-W private test set.

E. Experiment using COFW dataset

To evaluate the robustness to occlusion of the MSM method subject to various occluded face images, the COFW dataset is used which is regarded as a challenging dataset for existing state-of-the-art face alignment methods. In Table V, various methods including RCPR, TCDCN, HPM [46], CFSS, SHN, JMFA [33], AGCFN and LAB are compared. The MSM was trained on the 300-W dataset with a total of 3148 face training images. As shown in Table V, the MSM achieved the lowest inter-pupils NRMSE of 5.50% and the lowest inter-ocular NRMSE value of 3.90% with failure rate of 0%. These reflect the effectiveness of MSM in managing faces under heavy occlusion. The NRMSE values for SHN and JMFA are slightly higher than those of the MSM method. It should be noted that the training sets of both the SHN and the JMFA are much larger than that of the MSM in which the SHN and the JMFA include the 300-W and Menpo [47] training sets, for a total of 9360 face images, which is almost three times more images than that of the MSM.

Fig. 11 shows the CED curves which indicate the MSM outperforms other methods (including SAPM [60]) by a large margin on the COFW dataset. Example results obtained from COFW are given in Fig. 10.

F. Experiment using WFLW dataset

The landmark configurations of this dataset is different from above datasets, all images in WFLW dataset are annotated by
Fig. 10: MSM example outputs using COFW dataset subject to various occlusion, such as hands, glasses, food, and mask covering a wide range of faces.

### TABLE VI: NRMSE (%), failure rate (%) and AUC of face alignment results using WFLW dataset.

| Metric | Method       | Fullset | Pose | Expression | Illumination | Make-up | Occlusion | Blur |
|--------|--------------|---------|------|------------|--------------|---------|-----------|------|
| **NRMSE (%)** | SDM [22]    | 10.29  | 24.10 | 11.45      | 9.32         | 9.38    | 13.03     | 11.28|
|        | CFSS [49]   | 9.07   | 21.36 | 10.09      | 8.30         | 8.74    | 11.76     | 9.96 |
|        | DVLN [61]   | 6.08   | 11.54 | 6.78       | 5.73         | 5.98    | 7.33      | 6.88 |
|        | LAB [34]    | 5.27   | 10.24 | 5.51       | 5.23         | 5.15    | 6.79      | 6.32 |
|        | 3DDE [51]   | 4.68   | 8.62  | 5.21       | 4.65         | 4.60    | 5.77      | 5.41 |
|        | DeCaFA [55] | 4.62   | 8.11  | 4.65       | 4.41         | 4.63    | 5.74      | 5.38 |
|        | **MSM**     | **4.60** | **8.01** | **4.81** | **4.58**     | **4.47** | **5.85** | **5.28** |
| **Failure (%)** | SDM [22]    | 29.40  | 84.36 | 33.44      | 26.22        | 27.67   | 41.85     | 35.32|
|        | CFSS [49]   | 20.56  | 66.26 | 23.25      | 17.34        | 21.84   | 32.88     | 23.67|
|        | DVLN [61]   | 10.84  | 46.93 | 11.15      | 7.31         | 7.77    | 13.72     | 10.74|
|        | LAB [34]    | 7.56   | 28.83 | 6.37       | 6.73         | 7.77    | 13.72     | 10.74|
|        | 3DDE [51]   | 5.04   | 22.39 | 5.41       | 3.86         | 6.79    | 9.37      | 6.72 |
|        | DeCaFA [55] | 4.84   | 21.4  | 3.73       | **3.22**     | 6.15    | **9.26**  | 6.61 |
|        | **MSM**     | **4.28** | **16.87** | **2.87** | **3.72**     | **4.37** | **9.36** | **5.95** |
| **AUC** | SDM [22]    | 0.3002 | 0.0226 | 0.2293     | 0.3237       | 0.3125  | 0.2060    | 0.2398|
|        | CFSS [49]   | 0.3659 | 0.0632 | 0.3157     | 0.3854       | 0.3691  | 0.2688    | 0.3037|
|        | DVLN [61]   | 0.4551 | 0.1474 | 0.3889     | 0.4743       | 0.4494  | 0.3794    | 0.3973|
|        | LAB [34]    | 0.5323 | 0.2345 | 0.4951     | 0.5433       | 0.5394  | 0.4490    | 0.4630|
|        | 3DDE [51]   | 0.5544 | 0.2640 | 0.5175     | 0.5602       | 0.5536  | 0.4692    | 0.4957|
|        | DeCaFA [55] | 0.5630 | 0.2920 | 0.5460     | **0.5790**   | **0.5750** | **0.4850** | **0.4940** |
|        | **MSM**     | **0.5671** | **0.3091** | **0.5478** | **0.5725** | **0.5711** | **0.4849** | **0.5073** |

Fig. 11: CED curves of face alignment results using COFW dataset.

98-points manually. For comprehensive analysis of existing state-of-the-art methods, the dataset contains various type of challenge including large pose, illumination, blur, occlusion and excessive disturbing background, etc. Since WFLW is a newly released dataset, we compare the proposed method with a number of methods including ESR, SDM, CFSS, DVLN [61], LAB, 3DDE and DeCaFA [55]. We report the NRMSE (inter-ocular), failure rate and AUC on the test set and six subsets of WFLW. As shown in Table VI, the MSM method outperforms all other state-of-the-art methods in terms of the NRMSE, failure rate and AUC. An exception is for the case of an NRMSE value of 5.77% (occlusion subset) obtained by the 3DDE versus 5.85% obtained by the MSM. Note that the input images of 3DDE are cropped by ground-truth bounding box, which is much more beneficial to landmark localisation task. However, MSM still outperforms 3DDE using the provided bounding box in all other metrics. The MSM results using WFLW dataset are shown in Fig. 12.
Fig. 12: MSM example outputs using WFLW dataset subject to extremely challenge cases, such as illumination, large pose, occlusion and disturbing background, etc.

| Method       | Fullset | Pose | Expression | Illumination | Make-up | Occlusion | Blur |
|--------------|---------|------|------------|--------------|---------|-----------|------|
| Res-50       | 5.73    | 11.28| 6.13       | 5.65         | 5.80    | 6.98      | 6.51 |
| ST-GAN + Res-50 | 5.42   | 10.65| 6.00       | 5.31         | 5.39    | 6.57      | 6.23 |
| HG           | 5.41    | 10.03| 5.56       | 5.54         | 6.03    | 7.00      | 6.25 |
| ST-GAN + HG  | 4.81    | 8.49 | 5.09       | 4.75         | 4.70    | 6.16      | 5.51 |
| HG + SR      | 5.17    | 9.49 | 5.42       | 5.38         | 5.74    | 6.60      | 6.08 |
| ST-GAN + HG + SR | 4.60  | 8.01 | 4.81       | 4.58         | 4.47    | 5.85      | 5.28 |

**TABLE VIII:** Comparisons of NRMSE (%) and failure rate (%) using COFW dataset with different configurations.

| Method        | NRMSE (%) | Failure (%) |
|---------------|-----------|-------------|
| Res-50        | 4.76      | 4.54        |
| ST-GAN + Res-50 | 4.23    | 3.81        |
| HG            | 4.64      | 6.52        |
| ST-GAN + HG   | 4.34      | 5.23        |
| HG + SR       | 4.10      | 0.99        |
| ST-GAN + HG + SR | 3.95   | 0.99        |

**G. Experimental results on ablation study**

In this subsection the proposed method is evaluated by different configurations. The framework consists of several pivotal components including ST-GAN, stacked hourglass network and examplar-based face shape reconstruction. Their effectiveness are validated within the framework based on the COFW and WFLW datasets. To further evaluate the robustness of ST-GAN, a 50-layer residual network (Res-50) is introduced to verify whether the ST-GAN is effective to coordinate regression-based method. Since Res-50 requires input images size of $224 \times 224$, the size of the average pooling kernel in Res-50 is resized from 7 to 4, and the size of the network input is $128 \times 128$. The results of all ablation experiments use the inter-ocular distance as normalizing factor. Each proposed component was analyzed, i.e., with ST-GAN (labeled as ST-GAN), hourglass network (labeled as HG), and shape reconstruction (labeled as SR), by comparing their NRMSE and failure rates. Note that our baseline is HG, and ST-GAN+HG+SR represents the full MSM method.

Table VII and Table VIII show the NRMSE values and failure rates obtained by different configurations of our framework evaluated on the COFW and WFLW datasets. When combined with the ST-GAN, the Res-50 network reduces the NRMSE from 4.76% to 4.23%, and the hourglass network decrease
the NRMSE from 4.64% to 4.34%. This result demonstrates that the proposed ST-GAN method improved the performance of the face alignment task because STN can remove the translation, scale and rotation variation in each face, which can further reduce the variance in the regression target. Note that our method can effectively normalize face images to canonical poses and simultaneously remove unnecessary background. Compared with the baseline (HG) of our work, the innovations introduced in this paper exhibit a certain improvement for each subset of the WFLW dataset. These results demonstrate that in various difficult situations, the scoring scheme and face shape reconstruction method can be used to accurately locate difficult key points, not just in the case of occlusion. In Fig. 13, the CED curves show that ST-GAN+HG+SR which representing the full MSM method outperforms the other two configurations. Examples of the outputs obtained by the proposed ST-GAN on the WFLW dataset are shown in Fig. 14.

Finally, we discuss the setting of the threshold $K$ for distinguishing the reliability of landmarks. To this end, we performed a statistical analysis of the scores for each landmark of each sample on the COFW dataset, as shown in Fig. 15. As can be seen from the definition of the landmarks in Fig. 16, landmarks 1 to 17 in the contour of the face obtain significantly lower scores. This is because the features of the face contours are relatively simple. Conversely, features near

| Configuration | contour | facial features | NRMSE (%) |
|---------------|---------|----------------|-----------|
| 0.3           | 0.4     |                | 5.59      |
| 0.3           | 0.5     |                | 5.57      |
| 0.3           | 0.6     |                | 5.58      |
| 0.3           | 0.7     |                | 5.65      |
| 0.4           | 0.4     |                | 5.55      |
| 0.4           | 0.5     |                | 5.54      |
| 0.4           | 0.6     |                | 5.50      |
| 0.4           | 0.7     |                | 5.61      |
| 0.5           | 0.4     |                | 5.62      |
| 0.5           | 0.5     |                | 5.60      |
| 0.5           | 0.6     |                | 5.63      |
| 0.5           | 0.7     |                | 5.66      |
the facial features are significantly more discriminative, thus landmarks at these locations have higher scores. From the above analysis, we can draw a conclusion that it is unreliable to set the same threshold for all landmarks to distinguish the localization quality. The landmarks at the contour of the face should be set with lower thresholds, while the landmarks at the facial features of the face are in contrast. Therefore, we verified several different threshold configurations, as shown in Table IX. Finally, the setting for the threshold $K$ is: landmarks at the contour is 0.4, and landmarks at the facial features is 0.6.

V. Conclusion

In this paper a multistage model has been presented for robust face alignment. Our method leverages the best advantages of STNs, CNNs and exemplar-based shape constraints. Benefiting from the robust spatial transformation of the ST-GAN, the input image is warped to an alignment-friendly state. The stacked hourglass network provides accurate localization to landmarks that contain rich local information. The intensity of the heatmap is introduced to distinguish the aligned landmarks from missing landmarks, and the weight of each aligned landmark is determined simultaneously. Finally, with the help of these aligned landmarks, misaligned landmarks is refined by sparse shape constraints. A compact face shape dictionary learned by the K-means algorithm is used to improve the computational efficiency. Extensive experiments and ablation study have been conducted using challenging datasets (300-W, COFW and WFLW), the experimental results and analysis have demonstrated the effectiveness of the proposed multi-stage model as compared to other state-of-the-art methods. For portable and real-time applications, multiplierless neural networks [62]–[69] can be designed using back propagation [69] and other algorithms for implementing the multistage model.

Demos are posted on the website at http://101.37.150.44:8088/ism.asm.aspx.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 61372137, in part by the Natural Science Foundation of Anhui Province under Grant 1908085MF209 and Grant 1708085MF151, and in part by the Natural Science Foundation for the Higher Education Institutions of Anhui Province under Grant KJ2019A0036.

References

[1] F. Liu, Q. Zhao, X. Liu, and D. Zeng, “Joint face alignment and 3D face reconstruction with application to face recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, https://doi.org/10.1109/TPAMI.2018.28855995.
[2] C. Shan, S. Gong, and P. W. Mcowan, “Facial expression recognition based on local binary patterns: A comprehensive study,” Image and Vision Computing, vol. 27, no. 6, pp. 803–816, 2009.
[3] C. Cao, Q. Hou, and K. Zhou, “Displaced dynamic expression regression for real-time facial tracking and animation,” ACM Transactions on Graphics, vol. 33, no. 4, p. 43, 2014.
[4] A. Jourabloo and X. Liu, “Large-pose face alignment via CNN-based dense 3d model fitting,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4188–4196.
[5] E. Shelhamer, J. Long, and T. Darrell, “Fully convolutional networks for semantic segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 640–651, 2017.
[6] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in Proceedings of European Conference on Computer Vision, vol. 9905. Springer, 2016, pp. 2017–2025.
[7] X. P. Burgos-Artizzu, P. Perona, and P. Dollar, “Robust face landmark estimation under occlusion,” in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1513–1520.
[8] Y. Wu and Q. Ji, “Robust facial landmark detection under significant head poses and occlusion,” in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3658–3666.
[9] J. Xing, Z. Niu, J. Huang, W. Hu, X. Zhou, and S. Yan, “Towards robust and accurate multi-view and partially-occluded face alignment,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 987–1001, 2018.
[10] Q. Liu, J. Deng, J. Yang, G. Liu, and D. Tao, “Adaptive cascade regression model for robust face alignment,” IEEE Transactions on Image Processing, vol. 26, no. 2, pp. 797–807, 2017.
[11] S. Ren, X. Cao, Y. Wei, and J. Sun, “Face alignment via regressior local binary features,” IEEE Transactions on Image Processing, vol. 25, no. 3, pp. 1233–1245, 2016.
[12] J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou, “A deep regression architecture with two-stage re-initialization for high performance facial landmark detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3317–3326.
[13] J. Yang, Q. Liu, and K. Zhang, “Stacked hourglass network for robust facial landmark localization,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 79–87.
[14] X. Yan, H. Wang, Q. Wang, J. Song, and L. Tao, “Score-guided face alignment network under occlusions,” in Chinese Conference on Pattern Recognition and Computer Vision. Springer, 2018, pp. 195–206.
[15] M. Jaderberg, K. Simonyan, A. Zisserman et al., “Spatial transformer networks,” in Advances in Neural Information Processing Systems, 2015, pp. 2017–2025.
[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
[17] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” IEEE Transactions on Pattern Analysis and Machine Intelligence, no. 6, pp. 681–685, 2001.
[18] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, “Active shape models-their training and application,” Computer Vision and Image Understanding, vol. 61, no. 1, pp. 38–59, 1995.
[19] D. Cristinacce and T. F. Cootes, “Feature detection and tracking with constrained local models,” in Proceedings of British Machine Vision Conference. Citeseer, 2006, pp. 1–10.
[20] G. Tzimiropoulos and M. Pantic, “Gauss-newton deformable part models for face alignment in-the-wild,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1851–1858.
[21] T. F. Cootes, M. C. Ionita, C. Lindner, and P. Sauer, “Robust and accurate shape model fitting using random forest regression voting,” in Proceedings of European Conference on Computer Vision, vol. 3778. Springer Heidelberg, 2012, pp. 278–291.
[22] X. Xiong and F. De la Torre, “Supervised descent method and its applications to face alignment,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 532–539.
[23] X. Fan, R. Liu, Z. Luo, Y. Li, and Y. Feng, “Explicit shape regression with characteristic number for facial landmark localization,” IEEE Transactions on Multimedia, vol. 20, no. 3, pp. 567–579, 2018.
[24] J. Yan, Z. Lei, D. Yi, and S. Li, “Learn to combine multiple hypotheses for accurate face alignment,” in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2013, pp. 392–396.
[25] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, “300 faces in-the-wild challenge: The first facial landmark localization challenge,” in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2013, pp. 397–403.
[26] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[27] Y. Sun, X. Wang, and X. Tang, “Deep convolutional network cascade for facial point detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3476–3483.
[28] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “Learning deep representation for face alignment with auxiliary attributes,” IEEE Transactions on
