Dynamic Attention-controlled Cascaded Shape Regression Exploiting Training Data Augmentation and Fuzzy-set Sample Weighting

Zhen-Hua Feng[1] Josef Kittler[1] William Christmas[1] Patrik Huber[1] Xiao-Jun Wu[2]
[1] Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, UK
[2]School of IoT Engineering, Jiangnan University, Wuxi 214122, China
{z.feng, j.kittler, w.christmas, p.huber}@surrey.ac.uk, wu_xiaojun@jiangnan.edu.cn

Abstract

We present a new Cascaded Shape Regression (CSR) architecture, namely Dynamic Attention-Controlled CSR (DAC-CSR), for robust facial landmark detection on unconstrained faces. Our DAC-CSR divides facial landmark detection into three cascaded sub-tasks: face bounding box refinement, general CSR and attention-controlled CSR. The first two stages refine initial face bounding boxes and output intermediate facial landmarks. Then, an online dynamic model selection method is used to choose appropriate domain-specific CSRs for further landmark refinement. The key innovation of our DAC-CSR is the fault-tolerant mechanism, using fuzzy set sample weighting, for attention-controlled domain-specific model training. Moreover, we advocate data augmentation with a simple but effective 2D profile face generator, and context-aware feature extraction for better facial feature representation. Experimental results obtained on challenging datasets demonstrate the merits of our DAC-CSR over the state-of-the-art methods.

1. Introduction

Facial Landmark Detection (FLD), also known as face alignment, is a prerequisite for many automatic face analysis systems, *e.g.* face recognition [3, 33, 34], expression analysis [13, 14] and 2D-3D inverse rendering [1, 20, 21, 23, 28, 48]. Facial landmarks provide accurate shape information with semantic meaning, enabling geometric image normalisation and feature extraction for use in the remaining stages of a face processing pipeline. This is crucial for high-fidelity face image analysis. As the technology of FLD for constrained faces has already been well developed, the current trend is to address FLD for unconstrained faces in the presence of extreme variations in pose, expression, illumination and partial occlusion [2, 4, 24, 25, 30].

More recently, unconstrained FLD has seen huge progress owing to the state-of-the-art Cascaded Shape Regression (CSR) architecture [6, 12, 15, 29, 46]. The key to the success of CSR is to construct a strong regressor from a set of weak regressors arranged in a cascade. This architecture greatly improves the performance of FLD in terms of generalisation capacity and accuracy. However, in the light of very recent studies [35, 39, 42, 46, 47], the capacity of CSR appears to be saturating, especially for unconstrained faces with extreme appearance variations. For example, the FLD error of state-of-the-art CSR-based methods increases from around 3% (error in percent of the inter-ocular distance) on the Labelled Face Parts in the Wild (LFPW) [2] dataset to 6.5% on the more challenging Caltech Occluded Faces in the Wild (COFW) [4] dataset. This degradation has three main reasons: 1) The modelling capacity of the existing CSR architecture is limited. 2) CSR is sensitive to the positioning of face bounding boxes used for landmark initialisation. 3) The volume of available training data is insufficient. Can these limitations be overcome, especially for unconstrained faces exhibiting extreme appearance variations? We offer an encouraging answer by presenting a new Dynamic Attention-Controlled CSR (DAC-CSR) architecture with a dynamic domain selection mechanism and a novel training strategy which benefits from training data augmentation and fuzzy set training sample weighting.

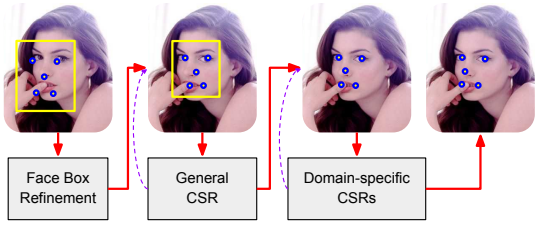

Figure 1. The pipeline of our proposed DAC-CSR.

Fig. 1 depicts a simplified overview of the proposed DAC-CSR architecture. Its innovation is in linking three types of regressor cascades performing in succession: 1) face bounding box refinement for better landmark initialisation, 2) an initial landmark update using a general CSR, and

3) a final landmark refinement by dynamically selecting an attention-controlled domain-specific CSR that is optimised to improve landmark location estimates. The new architecture decomposes the task at hand into three cascaded subtasks that are easier to handle.

In contrast to previous multi-view models, *e.g*. [39, 46], the key innovation of our DAC-CSR is its in-built fault-tolerant mechanism. The fault tolerance is achieved by means of an innovative training strategy for attention-controlled model training of the set of domain-specific CSRs performing the final shape update refinement. Rather than using samples from just a single domain, each domain-specific regressor cascade is trained using all the training samples. However, their influence is controlled by a domain-specific fuzzy membership function which weighs samples from the relevant domain more heavily than all the other training samples. An annealing schedule of domain-specific fuzzy membership functions progressively sharpens the relative weighting of in-domain and out-of-domain training samples in favour of the in-domain set for successive stages of each domain-specific cascade.

Each test sample progresses through the system of cascades. Prior to each of the domain-specific cascade stages, the domain of attention is selected dynamically based on the current shape estimate. The proposed training strategy guarantees that each domain-specific cascaded regressor can cope with out-of-domain test samples and is endowed with the capacity to update the shape in the correct direction even if the current domain has been selected subject to labelling error. This is the essence of error tolerance of the proposed system.

An important contributing factor to the promising performance of our DAC-CSR is training data augmentation. Our innovation here is to use a 2D face model for synthesising extreme profile face poses (out of plane rotation) with realistic background. Furthermore, we propose a novel context-aware feature extraction method to extract rich local facial features in the context of global face description.

The proposed framework has been evaluated on benchmarking databases using standard protocols. The results achieved on the database containing images with extreme poses (AFLW [24]) are significantly better than the state-of-the-art performance reported in the literature.

The paper is organised as follows. In the next section we present a brief review of related literature. The preliminaries of CSR are presented in Section 3. The proposed DAC-CSR is introduced in Section 4.1. The discussion of its training is confined to Section 4.2, which defines the domain-specific fuzzy membership functions and their annealing schedule. On-line dynamic domain selection is the subject of Section 4.3 and the proposed feature extraction scheme can be found in Section 4.4. Section 5 addresses the problem of training set augmentation. The experimental evaluation carried out and the results achieved are described in Section 6. The paper is drawn to conclusion in Section 7.

2. Related Work

Facial landmark detection can trace its history to the nineteen nineties. The representative FLD methods making the early milestones include Active Shape Model (ASM) [8], Active Appearance Model (AAM) [7] and Constrained Local Model (CLM) [10]. These algorithms and their extensions have achieved excellent FLD results in constrained scenarios [17]. As a result, the current trend is to develop a more robust FLD for unconstrained faces that are rich in appearance variations. The leading algorithms for unconstrained FLD are CSR-based approaches [6, 12, 15, 29, 46]. In contrast to the classical methods such as ASM, AAM and CLM that rely on a generative PCA-based shape model, CSR directly positions facial landmarks on their optimal locations based on image features. The shape update is achieved in a discriminative way by constructing a mapping function from robust shape-related local features to shape updates. The secret of the success of CSR is the architecture that cascades a set of weak regressors in series to form a strong regressor.

There have been a number of improvements to the performance of CSR-based FLD. One category of these improvements is to enhance some components of the existing CSR architecture. For example, the use of more robust shape-related local features, *e.g*. Scale-Invariant Feature Transform (SIFT) [38, 42, 43], Histogram of Oriented Gradients (HOG) [11, 15, 21, 40], Sparse Auto-Encoder (SAE) [16], Local Binary Features (LBF) [6, 29] and Convolutional Neural Networks (CNN-) based features [35, 37], has been suggested. Another example is to use more powerful regression methods as weak regressors in CSR, such as random forests [6, 29] and deep neural networks [32, 35, 37, 42, 43, 44]. Lately, 3D face models have been shown to positively impact FLD in challenging benchmarking datasets, especially in relation to faces with extreme poses [15, 26, 47].

Multi-view models: Another important approach is to adopt advanced CSR architectures, such as the use of multiple CSR models or constructing multi-view models. Feng *et al*. [16] constructed multiple CSR models by randomly selecting subsets from the original training set and fusing multiple outputs to produce the final FLD result. A similar idea has also been used in [41]. As an alternative, a multi-view FLD system employs a set of view-specific models that are able to achieve more accurate landmark detection for faces exhibiting specific views [9, 36, 46].

However, the use of multiple models or multi-view models is not without difficulties. One has to either estimate the view of a test image to select an appropriate model, or apply all view-specific models to a test image and then choose the

best result as the final output. For the former, implementing a model selection stage for unconstrained faces is hard in practice. An erroneously selected view-specific model may result in FLD failure. For the latter strategy, it is time-consuming to apply all the trained models to a test image. Also, the ranking of the outputs of different view-based models is non-trivial. In contrast to previous studies, our DAC-CSR addresses these issues by improving the fault-tolerance properties of a trained domain-specific model and by using an online dynamic model selection strategy.

Data augmentation: For a learning-based approach such as CSR, the availability of a large volume of training samples is essential. However, it is a tedious task to manually annotate facial landmarks for a large quantity of training data. To address this problem, data augmentation is widely used in CSR-based FLD. Traditional methods include random perturbation of initial landmarks, image flipping, image rotation, image blurring and adding noise to the original face images. However, none of these methods are able to inject new out-of-plane rotated faces to an existing training dataset. Recently, to augment a training set by samples with rich pose variations, the use of 3D face models has been suggested. For instance, Feng *et al*. [15, 23, 31] used a 3D morphable face model to synthesise a large number of 2D faces. However, the synthesised virtual faces lack realistic appearance variations especially in terms of background and expression changes. To mitigate this problem, they advocated a cascaded collaborative regression strategy to train a CSR from a mixture of real and synthesised faces. To generate realistic face images with pose variations, Zhu *et al*. fit a 3D shape model to 2D face images and generate profile face views from the reconstructed 3D shape information [47]. However, these 3D-based methods [15, 23, 47] require 3D face scans for model construction, which are expensive to capture. Also, it is difficult in practice to fit a 3D face model to 2D images. In this paper, we propose a simple but efficient 2D-based method to generate virtual faces with out-of-plane pose variations.

3. Cascaded Shape Regression (CSR)

Given an input face image $\mathbf{I}$ and the corresponding face bounding box $\mathbf{b} = [x_1, y_1, x_2, y_2]^T$ (coordinates of the upper left and lower right corners) of a detected face in the image, the goal of FLD is to output the face shape in the form of a vector, $\mathbf{s} = [x_1, y_1, ..., x_L, y_L]^T$, consisting of the coordinates of L pre-defined facial landmarks with semantic meaning such as eye centres and nose tip. To this end, we first initialise the face shape, $\mathbf{s}'$, by putting the mean shape into the bounding box. Then a trained CSR $\Phi = \{\phi(1), \phi(2), ..., \phi(N)\}$ is used to update the initial shape estimate, where Φ is a strong regressor consisting of N weak regressors. A weak regressor can be obtained using any regression method, such as linear regression, random forests and neural networks. In this paper, we use ridge regression as a weak regressor, *i.e*. $\phi = \{\mathbf{A}, \mathbf{e}\}$:

$$\phi : \delta\mathbf{s} = \mathbf{A} \cdot f(\mathbf{I}, \mathbf{s}') + \mathbf{e}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{2L \times N_f}$ is a projection matrix, N_f is the dimensionality of a shape-related feature vector extracted using $f(\mathbf{I}, \mathbf{s}')$, and $\mathbf{e} \in \mathbb{R}^{2L}$ is an offset. For the shape-related feature extraction, we apply local descriptors, *e.g*. HOG, to the neighbourhoods of all the facial landmarks of the current shape estimate and concatenate the extracted features into a long vector. The use of a weak regressor results in an update to the current shape estimate:

$$\mathbf{s}' \leftarrow \mathbf{s}' + \delta\mathbf{s}. \tag{2}$$

A trained CSR applies all the weak regressors in cascade to progressively update the shape estimate and obtain the final FLD result from an input image.

Given a training dataset $T = \{\mathbf{I}_i, \mathbf{b}_i, \mathbf{s}_i^*\}_{i=1}^I$ with I samples including face images, face bounding boxes and manually annotated facial landmarks, we first obtain the initial shape estimates, $\{\mathbf{s}_i'\}_{i=1}^I$, of all the training samples using the face bounding boxes provided. Then the shape update between the current shape estimate and ground-truth shape of the ith training sample can be calculated using $\delta\mathbf{s}_i^* = \mathbf{s}_i^* - \mathbf{s}_i'$. The first weak regressor is obtained using ridge regression by minimising the loss:

$$\arg\min_{\mathbf{A},\mathbf{e}} \sum_{i=1}^{I} ||\mathbf{A} \cdot f(\mathbf{I}_i, \mathbf{s}_i') + \mathbf{e} - \delta\mathbf{s}_i^*||_2^2 + \lambda||\mathbf{A}||_F^2, \tag{3}$$

where λ is the weight of the regularisation term. This is a typical least-square estimation problem with a closed-form solution [16, 38]. Last, this trained weak regressor is used to update the current shape estimates of all the training samples, which forms the training data for the second weak regressor. This procedure is repeated until all the N weak regressors are obtained.

4. Dynamic Attention-controlled CSR

4.1. Architecture

The architecture of the proposed DAC-CSR method has three cascaded stages: face bounding box refinement, general CSR and domain-specific CSR, as shown in Fig. 2. In fact, our DAC-CSR can be portrayed as a strong regressor $\Phi = \{\phi_b, \Phi_g, \Phi_d\}$, where ϕ_b is a weak regressor for face bounding box refinement, $\Phi_g = \{\phi_g(1), ..., \phi_g(N_g)\}$ is a classical CSR with N_g weak regressors, $\Phi_d = \{\Phi_d(1), ..., \Phi_d(M)\}$ is a strong regressor with M domain-specific CSRs and each of them has N_d weak regressors $\Phi_d(m) = \{\phi_d(m, 1), ..., \phi_d(m, N_d)\}$.

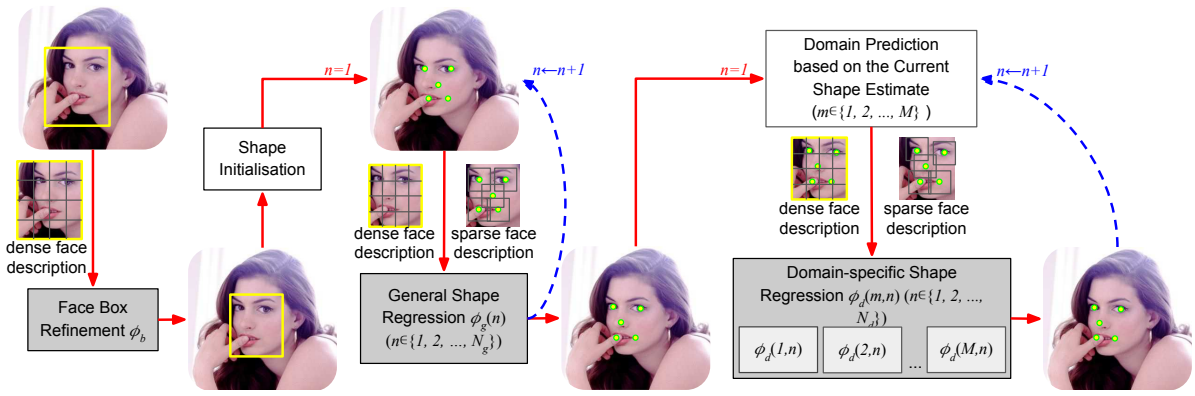

Figure 2. The proposed DAC-CSR has three stages in cascade: face bounding box refinement, general CSR and domain-specific CSR.

Face bounding box refinement: We define the weak regressor for the first step as $\phi_b = \{\mathbf{A}_b, \mathbf{e}_b\}$:

$$\phi_b : \delta\mathbf{b} = \mathbf{A}_b \cdot f_b(\mathbf{I}, \mathbf{b}) + \mathbf{e}_b, \tag{4}$$

where $f_b(\mathbf{I}, \mathbf{b})$ extracts dense local features from the image region inside the original face bounding box and $\delta\mathbf{b}$ is used to adjust the original bounding box.

The training of this weak regressor is the same as the procedure introduced in Section 3 for classical CSR. The only difference here is that we use face bounding box differences instead of shape differences for the regressor learning in Eq. (3). The ground-truth face bounding box for a training sample is computed by taking the minimum enclosing rectangle around the ground-truth face shape.

General CSR: The initial shape estimate, $\mathbf{s}'$, for general CSR is obtained by translating and scaling the mean shape so that it exactly fits into the refined bounding box, touching all four sides. Then the general CSR progressively updates the initial shape estimate, $\mathbf{s}' \leftarrow \mathbf{s}' + \delta\mathbf{s}$, using all the weak regressors in $\Phi_g = \{\phi_g(1), ..., \phi_g(N_g)\}$, as indicated in Algorithm 1. The nth weak regressor is defined as $\phi_g(n) = \{\mathbf{A}_g(n), \mathbf{e}_g(n)\}$:

$$\phi_g(n) : \delta\mathbf{s} = \mathbf{A}_g(n) \cdot f_c(\mathbf{I}, \mathbf{s}') + \mathbf{e}_g(n), \tag{5}$$

where $f_c(\mathbf{I}, \mathbf{s}')$ is a context-aware feature extraction function that combines both dense face description and shape-related sparse local features. The training of this stage is the same as the classical CSR introduced in Section 3.

Domain-specific CSR: Suppose this stage has M domain-specific CSRs corresponding to M sub-domains, each having N_d weak regressors. The nth weak regressor of the mth domain-specific CSR is defined as:

$$\phi_d(m, n) : \delta\mathbf{s} = \mathbf{A}_d(m, n) \cdot f_c(\mathbf{I}, \mathbf{s}') + \mathbf{e}_d(m, n), \tag{6}$$

where $m = 1, ..., M$, $N = 1, ..., N_d$. Given the current shape estimate $\mathbf{s}'$ output by the previous general CSR, a domain predictor is used to select a domain-specific CSR for the current shape update (Section 4.3). It should be noted that we use a dynamic domain selection strategy, which updates the label for the domain-specific model selection after each shape update, as shown in Algorithm 1. As a result of the proposed domain-specific CSR training described in Section 4.2, this mechanism makes our DAC-CSR tolerant to domain prediction errors.

input : image $\mathbf{I}$, face bounding box $\mathbf{b}$ and a trained DAC-CSR model $\Phi = \{\phi_b, \Phi_g, \Phi_d\}$
output: facial landmarks $\mathbf{s}'$

refine the face bounding box $\mathbf{b}$ using ϕ_b ;
estimate the current face shape, $\mathbf{s}'$, using the refined face bounding box ;
for $n \leftarrow 1$ **to** N_g **do**
  apply the nth general weak regressor $\phi_g(n)$ to update the current shape estimate;
end
for $n \leftarrow 1$ **to** N_d **do**
  predict the label (m) of the sub-domain of the current shape estimate using Eq. (11) ;
  apply the nth weak regressor $\phi_d(m, n)$ in the mth domain-specific CSR to update the current shape;
end

Algorithm 1: FLD using our DAC-CSR.

4.2. Offline Domain-specific CSR Training

Given a training dataset T with I samples, as introduced in Section 3, the first two stages, *i.e.* face bounding box refinement and general CSR, are trained directly using T. To train a domain-specific CSR, we first create M subsets $\{T_1, ..., T_M\}$ from the original training set, where $T_m \subseteq T$. To this end, we normalise all the current shape estimates, output by the previous general CSR, to the interval $[0, 1]$. Then PCA is used to obtain the first K shape eigenvectors. All the current shape estimates are projected to the

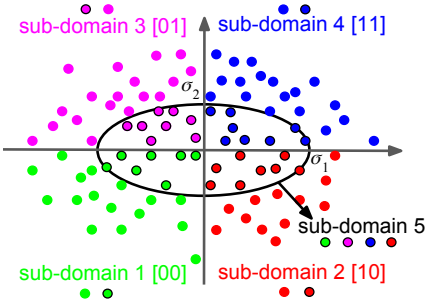

Figure 3. The proposed domain split strategy ($K = 2, \bar{c}_k = 0$).

K-dimensional subspace to obtain the projected coefficients $\{\mathbf{c}_i\}_{i=1}^{I}$, where $\mathbf{c}_i = [c_{i,1}, ..., c_{i,K}]^T$. Then the original domain is partitioned into $M = 2^K + 1$ overlapping sub-domains, as demonstrated in Fig. 3 for $K = 2$. For the Mth sub-domain, it includes the training samples satisfying $\sum_{k=1}^{K} \frac{(c_{i,k}-\bar{c}_k)^2}{(\sigma(k))^2} \leq 1$, where $\bar{c}_k$ and $\sigma(k)$ are the mean and standard deviation of the kth element of the coefficient vectors. For other sub-domains, each includes the training samples in a specific region of a K-dimensional coordinate system. To be more specific, for each coefficient $\mathbf{c}_i$, a sub-domain membership word $g(\mathbf{c}_i)$ is generated by:

$$g(\mathbf{c}_i) = 1 + \sum_{k=1}^{K} b_c(c_{i,k}) 2^{k-1}, \tag{7}$$

where $b_c(c_{i,k})$ is a coding function that converts the kth element in a coefficient vector to a bit:

$$b_c(c_{i,k}) = \begin{cases} 1 & \text{if } c_{i,k} \geq \bar{c}(k) \\ 0 & \text{otherwise} \end{cases}. \tag{8}$$

Then the mth sub-domain, $1 < m < 2^K$, includes the training samples with their membership words $g(\mathbf{c}_i) = m$. Our domain split strategy results in M sub-domains with overlapping boundaries. This is different from previous studies using multi-view models such as [39, 46], in which the intersection of any two different subsets is empty, *i.e.* $T_i \cap T_j = \emptyset, \forall i \neq j$.

The advantage of our domain split strategy is that it improves the fault-tolerance ability of each trained domain-attention model, because of the overlap of two different sub-domains. For a test sample, a domain predictor may output an inaccurate label for model selection due to the rough shape estimate provided from the previous general CSR. But, the inaccurately selected domain-specific model is still able to refine the current shape estimate. To further improve this refinement capacity, we propose a fuzzy training strategy. For each domain-specific CSR, we use all the training samples from the original training set to train a specific regressor, but weight more heavily the training samples of the specific domain by increasing their fuzzy set membership values in the objective function. More specifically, to train the nth weak regressor of the mth domain-specific CSR, the objective function is defined as:

$$\arg\min_{\mathbf{A_d}, \mathbf{e_d}} \sum_{i=1}^{I} w_i ||\mathbf{A_d} \cdot f_c(\mathbf{I}_i, \mathbf{s}_i') + \mathbf{e_d} - \delta \mathbf{s}_i^*||_2^2 + \lambda ||\mathbf{A}_d||_F^2, \tag{9}$$

where w_i is a fuzzy set membership value defined by:

$$w_i = \begin{cases} 1 - h(n) & \text{if } \{\mathbf{I}_i, \mathbf{b}_i, \mathbf{s}_i^*\} \in T_m \\ h(n) & \text{otherwise} \end{cases}, \tag{10}$$

where $h(n)$ is a decreasing function which progressively reduces the weights of the training samples not belonging to the mth sub-domain and increases the weights of the training samples of the mth sub-domain. This is a standard weighted least-square estimation problem with a closed-form solution. It should be noted that our fuzzy domain-specific model learns a weak regressor that is able to refine a face shape estimate from any sub-domain, and with better capacity to refine face shapes from a specific domain. This capability is exhibited even when using a domain split strategy without overlap.

4.3. Dynamic Domain Selection in Testing

Given a new test image with a detected face bounding box, the trained DAC-CSR model $\Phi = \{\phi_b, \Phi_g, \Phi_d\}$ first applies the face bounding box refiner ϕ_b and general CSR Φ_g to obtain the intermediate face shape estimate $\mathbf{s}'$. Then a specific domain-attention weak regressor is selected to further update the current shape estimate.

To select an appropriate weak regressor, the current shape estimate $\mathbf{s}'$ is projected into the PCA space learned at training time to obtain the coefficient vector $\mathbf{c}$, and the label of the sub-domain is obtained using:

$$p(\mathbf{c}) = \begin{cases} 2^K + 1 & \text{if } \sum_{k=1}^{K} \frac{(c_{i,k}-\bar{c}_k)^2}{(\sigma(k))^2} \leq 1 \\ g(\mathbf{c}) & \text{otherwise} \end{cases}. \tag{11}$$

Note that, here, the sub-domains are not overlapped. This is different from the domain split strategy used in the training stage. However, this domain prediction function is only based on the current shape information and may provide inaccurate labels for model selection. To address this issue and further improve the fault-tolerance capacity of our DAC-CSR, a dynamic domain selection strategy is used.

As discussed in the last section, a trained domain-specific CSR is able to improve the current shape estimate even if selected in error by the domain prediction mechanism. Hence the updated shape estimate produced by the nth weak regressor can be a basis for selecting a more appropriate domain in the next step of the shape updating process. We re-run the domain prediction before performing the next weak regressor and choose the $(n+1)$st weak regressor of a newly selected domain-specific model for cur-

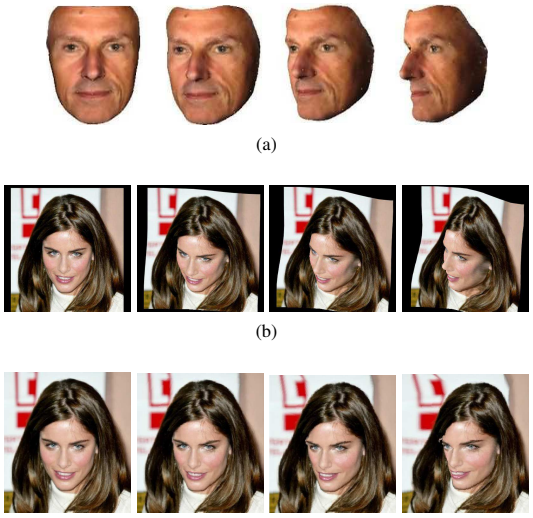

Figure 4. A comparison of synthesised 2D faces using (a) a 3D morphable model [15], (b) 3D-based face profiling [47], and (c) our 2D-based method.

rent shape update, as summarised in Algorithm 1. This dynamic model selection strategy is repeated after each shape update in our domain-specific CSR.

4.4. Context-aware Feature Extraction

Feature extraction is crucial for constructing a robust mapping from feature space to shape updates. In classical CSR-based approaches, shape-related local features are created by concatenating all the extracted local features around each landmark into a long vector. Although this sparse shape-related feature extraction method provides a good description of the texture information of different facial parts, it does not offer a good representation of the contextual information of faces. In our DAC-CSR, we use a context-aware feature extraction method. To be more specific, we use both a dense local description of the whole face region and sparse shape-related local features for weak regressor training (Fig. 3). Note that, for the first bounding box refinement step, we use only the dense local features.

5. 2D Profile Face Generation

For a learning-based approach, a large number of annotated face images are crucial for training. As discussed in Section 2, traditional data augmentation methods are not able to inject new out-of-plane pose variations, and the use of 3D face models is very expensive. To mitigate this issue, we propose a simple 2D-based method that can generate virtual faces with out-of-plane pose variations. A comparison between our proposed 2D-based profile face generator and two 3D-based methods [15, 47] is shown in Fig. 4.

To warp a face image to another pose, we first build

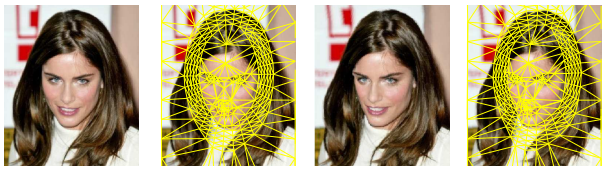

Figure 5. The mesh generated for 2D image warpping.

a PCA-based shape model that is equivalent to the shape model used in ASM [8] and AAM [7, 27]. Then we choose the corresponding shape eigenvector controlling yaw rotations (usually the first one) to change the pose of the current face shape. To this end, we first calculate the coefficient of the shape of a face image projected by the selected shape eigenvector. A new face shape with pose variations is generated by adjusting the projected coefficient. The 2D shape model used is constructed using a face dataset rich in pose variations. Note, we only generate pose-varying face shapes with the same rotation direction of the original shape, *i.e.* left or right. Then we expand the face shape with more external facial landmarks and compute a 2D mesh of the original and new shapes using Delaunay triangulation, as shown in Fig. 5. Last, a piece-wise affine warp is used to map the texture from the original face shape to a new one [27]. Moreover, the synthesised faces can be flipped about their vertical axis to obtain more faces with pose variations in the other direction (right or left), which is similar to [47].

6. Experimental Results

6.1. Datasets and Implementation Details

Datasets: In our experiments, we use two challenging face datasets, including the Annotated Facial Landmarks in the Wild (AFLW) dataset [24] and the Caltech Occluded Faces in the Wild (COFW) dataset [4] to evaluate the performance of our DAC-CSR architecture.

The AFLW dataset has 25993 unconstrained faces with large-scale pose variations up to $\pm 90^{\circ}$. Each AFLW face image has up to 21 landmarks of visible facial parts. AFLW does not have a standard protocol for FLD evaluation; hence we follow the protocol used in Cascaded Compositional Learning (CCL) [46]. This is the first work to use the whole AFLW dataset to benchmark an FLD algorithm. It reports the currently best results on AFLW. CCL used 24386 images from AFLW and manually annotated all the missing landmarks in the original dataset. The annotation system opted for 19 landmarks per image without the two ear landmarks (ID-13 and ID-17). CCL has two protocols: AFLW-full and AFLW-frontal, as shown in Table 1. AFLW-full splits the 24386 images into 20000/4386 for training/testing. The AFLW-frontal protocol selects 1165 [1]

[1] In our experiments, 1314 frontal faces were selected using the list provided by [46].

Table 1. A summary of the evaluation protocols used in our experiments

Protocol	Training Set	Test Set	# Landmarks	Normalisation	Setting
AFLW-full	20000 from AFLW	4386 from AFLW	19	face size	CCL [46]
AFLW-frontal	20000 from AFLW	1165 from AFLW	19	face size	CCL [46]
COFW	1345 from COFW	507 from COFW	29	eye distance	standard [4]

frontal images from the 4386 test images to evaluate an FLD algorithm on frontal faces.

The COFW dataset has 1345 training and 507 test images, which are all unconstrained faces. Each COFW face has 29 manually annotated landmarks. COFW is a challenging benchmark containing major occlusions.

Implementation Details: In our experiments, we only used one weak regressor for face bounding box refinement. The numbers of weak regressors for general CSR and domain-specific CSR were set to 2 and 3 respectively. We set the number of sub-domains to $M = 5$ using 2 PCA shape coefficients, *i.e.* $K = 2$. The value of the regularisation term in the ridge regression training was assigned to $\lambda = 10000$, and the decreasing schedule controlling fuzzy membership values was set to $h(n) = (0.3, 0.2, 0.1)$ for $n = (1, 2, 3)$. To extract a dense face description, we resized the face region to 100×100 and extracted HOG features using a cell size of 10 and block size of 2. To extract sparse shape-related local features, we computed the HOG descriptor in the neighbourhood of each facial landmark. The radius was set to $1/7$ of the maximum of the height and width of the current shape estimate. Each local image patch was resized to 30×30 and the cell size was set to 10. In addition, the central 15×15 image patch was used to extract multi-scale HOG features using a cell size of 5.

To augment training data, we applied our 2D-based method to generate virtual face images with new poses. Each training image in COFW was augmented using 9 new poses. For AFLW, we only synthesised new faces for semi-frontal training images. We also flipped all the training images about the vertical axis, added Gaussian blur with $\sigma = 1$ pixel and performed random perturbations of the initial face bounding boxes.

6.2. Evaluation on AFLW

The Cumulative Error Distribution (CED) curve of our DAC-CSR using the AFLW-full protocol is shown in Fig. 6. The error was calculated using the Euclidean distance between the detected and ground-truth landmarks, normalised by face size [46]. Our DAC-CSR achieves much better results on the AFLW-full protocol than the current best result reported for CCL [46].

Table 2 compares our DAC-CSR with state-of-the-art methods on AFLW using both the AFLW-full and AFLW-frontal protocols. The results obtained with our DAC-CSR show the best normalised average error on both the full test set and the frontal face subset protocols.

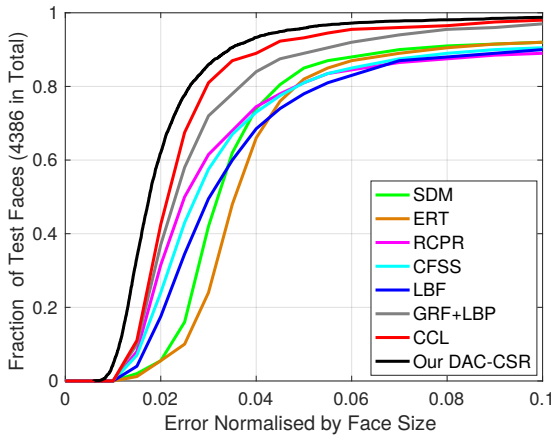

Figure 6. A CED curve comparison of our DAC-CSR with state-of-the-art methods, including SDM [38], ERT [22], RCPR [4], CFSS [45], LBF [29], LBF + GRF [19] and CCL [46], on the AFLW dataset (better viewed in colour). In this experiment, 20000 images were used for training and 4386 images were used for testing, following the **AFLW-full** protocol in [46].

Table 2. A comparison of our DAC-CSR with state-of-the-art methods on **AFLW**, measured in terms of the average error, normalised by face size. The protocol is the same as in [46].

Method	AFLW-full	AFLW-frontal
SDM [38]	4.05%	2.94%
RCPR [4]	3.73%	2.87%
ERT [22]	4.35%	2.75%
LBF [29]	4.25%	2.74%
LBF + GRF [19]	3.15%	N.A.
CFSS [45]	3.92%	2.68%
CCL [46]	2.72%	2.17%
Our DAC-CSR	**2.27%**	**1.81%**

6.3. Evaluation on COFW

6.3.1 Comparison to State-of-the-art

The CED curves of our DAC-CSR and a set of state-of-the-art methods on the COFW dataset are shown in Fig. 7. In addition, a more detailed comparison is presented in Table 3, reporting the average error, failure rate and speed. The failure rate is defined by the percentage of test images with more than 10% detection error.

Our DAC-CSR achieves competitive results in accuracy

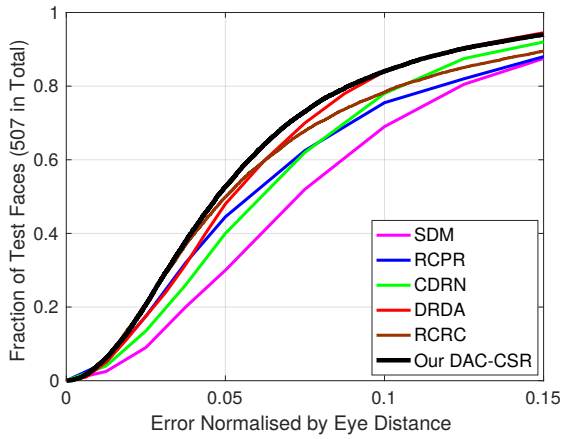

Figure 7. A comparison between our DAC-CSR and state-of-the-art methods, including SDM [38], RCPR [4], RCRC [16], CDRN [42] and DRDA [42], on **COFW**.

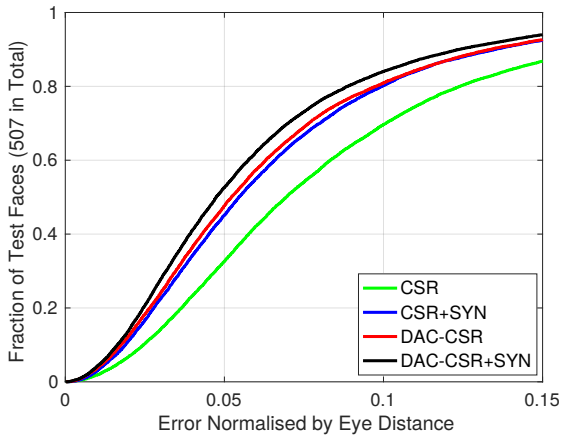

Figure 8. A self-evaluation of our proposed DAC-CSR on **COFW**. The meaning of each term is introduced in Section 6.3.2.

Table 3. Comparison on **COFW**. The error was measured on 29 landmarks and normalised by the inter-ocular distance.

Method	Error	Failure	Speed (FPS)
ESR [5]	11.2%	36%	4
RCPR [4]	8.5%	20%	3
HPM [18]	7.5%	13%	0.03
RCRC [16]	7.3%	12%	22
CCR [15]	7.03%	10.9%	**69**
DRDA [42]	6.46%	6%	N.A.
RAR [37]	6.03%	**4.14%**	4 (GPU)
Our DAC-CSR	**6.03%**	4.73%	10

compared to the two cutting-edge deep-neural-network-based algorithms, DRDA [42] and RAR [37]. In addition, the speed of our DAC-CSR on an Intel i7-4790 CPU is up to 10 FPS, which is faster than RAR with GPU acceleration (NVIDIA Titan Z). As the current bottleneck for unconstrained FLD is not the speed, *e.g.* LBF can perform FLD at up to 3000 FPS, the key aim of our DAC-CSR is to provide a more robust FLD algorithm for faces with extreme appearance variations, as exhibited in the AFLW-full evaluation.

6.3.2 Self Evaluation

In this part, we investigate the contributions of the proposed DAC-CSR architecture and our 2D-based data augmentation method to the accuracy of FLD on COFW. To this end, we compare the classical CSR method trained on the original training set (CSR) with the classical CSR trained on the augmented dataset using faces synthesised by our 2D-based face generation method (CSR+SYN), our DAC-CSR trained on the original dataset (DAC-CSR) and our DAC-CSR trained on the augmented dataset (DAC-CSR+SYN). The CED curves of these settings are shown in Fig. 8.

In fact, the architecture of classical CSR is the same as SDM [38]. They also have similar CED curves (comparing Fig. 7 with Fig. 8). As indicated by Fig. 8, the new DAC-CSR architecture trained on the original dataset performs better than CSR with our 2D-based data augmentation method (DAC-CSR vs CSR+SYN). However, the best result is achieved when the new DAC-CSR architecture is used jointly with our 2D-based data augmentation method.

7. Conclusion

We have presented a new DAC-CSR architecture for robust FLD in unconstrained faces. The proposed method achieved superior FLD results on the challenging AFLW dataset and delivered competitive performance on the COFW dataset. This is due to the proposed versatile fault-tolerant mechanism using fuzzy domain-specific model training and the online dynamic model selection strategy. In addition, a simple but effective data augmentation method based on 2D face synthesis was proposed. Compared with the classical CSR method, both the new DAC-CSR architecture and the 2D-based data augmentation method proved beneficial for the FLD performance on unconstrained faces.

We believe that our contributions can be further extended, *e.g.* using deep-neural-network-based approaches. We leave for future work the exploration of methods that combine our DAC-CSR architecture and data augmentation method with other FLD algorithms.

Acknowledgements

This work was supported in part by the EPSRC Programme Grant 'FACER2VM' (EP/N007743/1), the National Natural Science Foundation of China (61373055, 61672265) and the Natural Science Foundation of Jiangsu Province (BK20140419, BK20161135).

References

[1] O. Aldrian and W. A. P. Smith. Inverse rendering of faces with a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(5):1080–1093, 2013.

[2] P. N. Belhumeur, D. W. Jacobs, D. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 545–552, 2011.

[3] J. R. Beveridge, H. Zhang, B. A. Draper, P. J. Flynn, Z.-H. Feng, P. Huber, J. Kittler, Z. Huang, S. Li, Y. Li, et al. Report on the FG 2015 video person recognition evaluation. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–8. IEEE, 2015.

[4] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *International Conference on Computer Vision*, 2013.

[5] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by Explicit Shape Regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2887–2894. IEEE, 2012.

[6] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014.

[7] T. F. Cootes, G. Edwards, and C. J. Taylor. Active appearance models. In *European Conference on Computer Vision*, volume 1407, pages 484–498, 1998.

[8] T. F. Cootes and C. J. Taylor. Active shape models - 'smart snakes'. In *British Machince Vision Conference*, pages 266–275, 1992.

[9] T. F. Cootes, G. V. Wheeler, K. N. Walker, and C. J. Taylor. View-based active appearance models. *Image and Vision Computing*, 20(9):657–664, 2002.

[10] D. Cristinacce and T. F. Cootes. Feature Detection and Tracking with Constrained Local Models. In *British Mahine Vision Conference*, volume 3, pages 929–938, 2006.

[11] J. Deng, Q. Liu, J. Yang, and D. Tao. M3csr: Multi-view, multi-scale and multi-component cascade shape regression. *Image and Vision Computing*, 47:19–26, 2016.

[12] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1078–1085. IEEE, 2010.

[13] S. Eleftheriadis, O. Rudovic, M. P. Deisenroth, and M. Pantic. Variational gaussian process auto-encoder for ordinal prediction of facial action units. In *Asian Conference on Computer Vision*, Taipei, Taiwan. Oral, November 2016.

[14] S. Eleftheriadis, O. Rudovic, and M. Pantic. Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. *IEEE transactions on Image Processing*, 24(1):189–204, 2015.

[15] Z.-H. Feng, G. Hu, J. Kittler, W. Christmas, and X.-J. Wu. Cascaded collaborative regression for robust facial landmark detection trained using a mixture of synthetic and real images with dynamic weighting. *IEEE Transactions on Image Processing*, 24(11):3425–3440, 2015.

[16] Z.-H. Feng, P. Huber, J. Kittler, W. Christmas, and X. Wu. Random Cascaded-Regression Copse for Robust Facial Landmark Detection. *IEEE Signal Processing Letters*, 22(1):76–80, Jan 2015.

[17] Z.-H. Feng, J. Kittler, W. Christmas, X.-J. Wu, and S. Pfeiffer. Automatic face annotation by multilinear AAM with missing values. In *International Conference on Pattern Recognition (ICPR)*, pages 2586–2589. IEEE, 2012.

[18] G. Ghiasi and C. C. Fowlkes. Occlusion Coherence: Localizing Occluded Faces with a Hierarchical Deformable Part Model. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2014.

[19] K. Hara and R. Chellappa. Growing regression forests by classification: Applications to object pose estimation. In *European Conference on Computer Vision, (ECCV)*, pages 552–567. Springer, 2014.

[20] G. Hu, F. Yan, J. Kittler, W. Christmas, C.-H. Chan, Z.-H. Feng, and P. Huber. Efficient 3D Morphable Face Model Fitting. *Pattern Recognition*, 67:366–379, 2017.

[21] P. Huber, Z.-H. Feng, W. Christmas, J. Kittler, and M. Rätsch. Fitting 3D Morphable Face Models using local features. In *IEEE International Conference on Image Processing (ICIP)*, pages 1195–1199. IEEE, 2015.

[22] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1867–1874, 2014.

[23] J. Kittler, P. Huber, Z.-H. Feng, G. Hu, and W. Christmas. 3D Morphable Face Models and Their Applications. In *International Conference on Articulated Motion and Deformable Objects*, pages 185–206. Springer, 2016.

[24] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. In *First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[25] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *European Conference on Computer Vision*, pages 679–692. Springer, 2012.

[26] F. Liu, D. Zeng, Q. Zhao, and X. Liu. Joint face alignment and 3d face reconstruction. In *European Conference on Computer Vision (ECCV)*, pages 545–560. Springer, 2016.

[27] I. Matthews and S. Baker. Active Appearance Models Revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[28] M. Piotraschke and V. Blanz. Automated 3d face reconstruction from multiple images using quality measures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[29] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1685–1692, 2014.

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

[31] X. Song, Z.-H. Feng, G. Hu, J. Kittler, W. Christmas, and X.-J. Wu. Dictionary Integration using 3D Morphable Face

Models for Pose-invariant Collaborative-representation-based Classification. *arXiv preprint arXiv:1611.00284*, 2016.

[32] Y. Sun, X. Wang, and X. Tang. Deep Convolutional Network Cascade for Facial Point Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3476–3483, 2013.

[33] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[34] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1701–1708, 2014.

[35] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou. Mnemonic Descent Method: A Recurrent Process Applied for End-To-End Face Alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[36] O. Tuzel, T. K. Marks, and S. Tambe. Robust face alignment using a mixture of invariant experts. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *European Conference on Computer Vision (ECCV)*, pages 825–841, Cham, 2016. Springer International Publishing.

[37] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *European Conference on Computer Vision (ECCV)*, 2016.

[38] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 532–539, 2013.

[39] X. Xiong and F. De la Torre. Global supervised descent method. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2664–2673, 2015.

[40] J. Yan, Z. Lei, D. Yi, and S. Z. Li. Learn to Combine Multiple Hypotheses for Accurate Face Alignment. In *International Conference of Computer Vision - Workshops*, 2013.

[41] H. Yang, X. Jia, I. Patras, and K.-P. Chan. Random Subspace Supervised Descent Method for Regression Problems in Computer Vision. *Signal Processing Letters, IEEE*, 22(10):1816–1820, 2015.

[42] J. Zhang, M. Kan, S. Shan, and X. Chen. Occlusion-Free Face Alignment: Deep Regression Networks Coupled With De-Corrupt AutoEncoders. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[43] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-Fine Auto-Encoder Networks (CFAN) for Real-Time Face Alignment. In *European Conference on Computer Vision*, volume 8690, pages 1–16. Springer International Publishing, 2014.

[44] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108. Springer, 2014.

[45] S. Zhu, C. Li, C. Change Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4998–5006, 2015.

[46] S. Zhu, C. Li, C.-C. Loy, and X. Tang. Unconstrained Face Alignment via Cascaded Compositional Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[47] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face Alignment Across Large Poses: A 3D Solution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[48] X. Zhu, J. Yan, D. Yi, Z. Lei, and S. Z. Li. Discriminative 3D morphable model fitting. In *IEEE International Conference and Workshops onAutomatic Face and Gesture Recognition*, volume 1, pages 1–8, 2015.