Wing Loss for Robust Facial Landmark Localisation with Convolutional Neural Networks

Zhen-Hua Feng[1] Josef Kittler[1] Muhammad Awais[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, m.a.rana}@surrey.ac.uk, patrikhuber@gmail.com, wu_xiaojun@jiangnan.edu.cn

Abstract

We present a new loss function, namely Wing loss, for robust facial landmark localisation with Convolutional Neural Networks (CNNs). We first compare and analyse different loss functions including L2, L1 and smooth L1. The analysis of these loss functions suggests that, for the training of a CNN-based localisation model, more attention should be paid to small and medium range errors. To this end, we design a piece-wise loss function. The new loss amplifies the impact of errors from the interval (-w, w) by switching from L1 loss to a modified logarithm function.

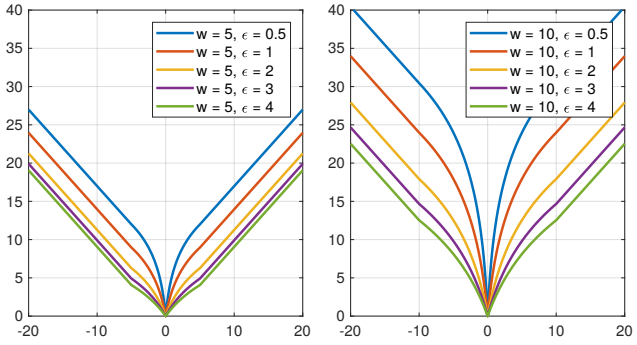
To address the problem of under-representation of samples with large out-of-plane head rotations in the training set, we propose a simple but effective boosting strategy, referred to as pose-based data balancing. In particular, we deal with the data imbalance problem by duplicating the minority training samples and perturbing them by injecting random image rotation, bounding box translation and other data augmentation approaches. Last, the proposed approach is extended to create a two-stage framework for robust facial landmark localisation. The experimental results obtained on AFLW and 300W demonstrate the merits of the Wing loss function, and prove the superiority of the proposed method over the state-of-the-art approaches.

1. Introduction

Facial landmark localisation, or face alignment, aims at finding the coordinates of a set of pre-defined key points for 2D face images. A facial landmark usually has specific semantic meaning, *e.g.* nose tip or eye centre, which provides rich geometric information for other face analysis tasks such as face recognition [55, 40, 37, 67], emotion estimation [69, 16, 57, 35] and 3D face reconstruction [15, 31, 26, 25, 48, 33, 19].

Thanks to the successive developments in this area of research during the past decades, we are able to perform very accurate facial landmark localisation in constrained scenarios, even using traditional approaches such as Active Shape Model (ASM) [7], Active Appearance Model (AAM) [8] and Constrained Local Model (CLM) [11]. The existing challenge is to achieve robust and accurate landmark localisation of unconstrained faces that are impacted by a variety of appearance variations, *e.g.* in pose, expression, illumination, image blurring and occlusion. To this end, cascaded-regression-based approaches have been widely used, in which a set of weak regressors are cascaded to form a strong regressor [13, 63, 6, 18, 60, 58, 20]. However, the capability of cascaded regression is nearly saturated due to its shallow structure. After cascading more than four or five weak regressors, the performance of cascaded regression is hard to improve further [52, 17]. More recently, deep neural networks have been put forward as a more powerful alternative in a wide range of computer vision and pattern recognition tasks, including facial landmark localisation [53, 72, 70, 39, 66, 59, 44].

To perform robust facial landmark localisation us-



(a) $w = 5$ (b) $w = 10$

Figure 1. Our Wing loss function (Eq. 5) plotted with different parameter settings, where w limits the range of the non-linear part and ϵ controls the curvature. By design, we amplify the impact of the samples with small and medium range errors to the network training.

ing deep neural networks, different network types have been explored, such as the Convolutional Neural Network (CNN) [53], Auto-Encoder Network [71] and Recurrent Neural Network (RNN) [56, 62]. In addition, different network architectures have been extensively studied during the recent years along with the development of deep neural networks in other AI applications. For example, the Fully Convolutional Network (FCN) [36] and hourglass network with residual blocks have been found very effective [43, 66, 12].

One crucial aspect of deep learning is to define a loss function leading to better-learnt representation from underlying data. However, this aspect of the design seems to be little investigated by the facial landmark localisation community. To the best of our knowledge, most existing facial landmark localisation approaches using deep learning are based on the L2 loss. However, the L2 loss function is sensitive to outliers, which has been noted in connection with the bounding box regression problem in the well-known Fast R-CNN algorithm [22]. Rashid *et al*. also notice this issue and use the smooth L1 loss instead of L2 [45]. To further address the issue, we propose a new loss function, namely Wing loss (Fig. 1), for robust facial landmark localisation. The main contributions of our work include:

- presenting a systematic analysis of different loss functions that could be used for regression-based facial landmark localisation with CNNs, which to our best knowledge is the first such study carried out in connection with the landmark localisation problem. We empirically and theoretically compare L1, L2 and smooth L1 loss functions and find that L1 and smooth L1 perform much better than the widely used L2 loss.

- a novel loss function, namely the Wing loss, which is designed to improve the deep neural network training capability for small and medium range errors.

- a data augmentation strategy, *i.e*. pose-based data balancing, that compensates the low frequency of occurrence of samples with large out-of-plane head rotations in the training set.

- a two-stage facial landmark localisation framework for performance boosting.

The paper is organised as follows. Section 2 presents a brief review of the related literature. The regression-based facial landmarking problem with CNNs is formulated in Section 3. The properties of common loss functions (L1 and L2) are discussed in Section 4 which also motivate the introduction of the novel Wing loss function. The pose-based data balancing strategy is the subject of Section 5. The two-stage localisation framework is proposed in Section 6. The advocated approach is validated experimentally in Section 7 and the paper is drawn to conclusion in Section 8.

2. Related work

Network Architectures: Most deep-learning-based facial landmark localisation approaches are regression-based. For such a task, the most straightforward way is to use a CNN model with regression output layers [53, 45]. The input for a regression CNN is usually an image patch enclosing the whole face region and the output is a vector consisting of the 2D coordinates of facial landmarks. Besides the classical CNN architecture, newly developed CNN systems have also been used for facial landmark localisation and shown promising results, *e.g*. FCN [36] and the hourglass network [43, 66, 12, 3, 4]. Different from traditional CNN-based approaches, FCN and hourglass network output a heat map for each landmark. These heat maps are of the same size as the input image. The value of a pixel in a heat map indicates the probability that its location is the predicted position of the corresponding landmark. To reduce false alarms of a generated 2D sparse heat map, Wu *et al*. propose a distance-aware softmax function that facilitates the training of their dual-path network [61].

Thanks to the extensive studies of different deep neural networks and their use cases in unconstrained facial landmark localisation, the development of the area has been greatly promoted. However, the current research lacks a systematic analysis on the use of different loss functions. In this paper, we close this gap and design a new loss function for CNN-based facial landmark localisation.

Dealing with Pose Variations: Extreme pose variations bring many difficulties to unconstrained facial landmark localisation. To mitigate this issue, different strategies have been explored. The *first* one is to use multi-view models. There is a long history of the use of multi-view models in landmark localisation, from the earlier studies on ASM [47] and AAM [10] to recent work on cascaded-regression-based [64, 75, 21] and deep-learning-based approaches [12]. For example, Feng *et al*. train multi-view cascaded regression models using a fuzzy membership weighting strategy, which, interestingly, outperforms even some deep-learning-based approaches [21]. The *second* strategy, which has become very popular in recent years, is to use 3D face models [76, 28, 2, 38, 29]. By recovering the 3D shape and estimating the pose of a given input 2D face image, the issue of extreme pose variations can be alleviated to a great extent. In addition, 3D face models have also been widely used to synthesise additional 2D face images with pose variations for the training of a pose-invariant system [41, 17, 76]. *Last*, multi-task learning has been adopted to address the difficulties posed by image degradation, including pose variations. For example, face attribute estimation, pose estimation or 3D face reconstruction can jointly be trained with facial landmark localisation [72, 65, 44]. The collaboration of different tasks in a multi-task learning framework can boost the performance of individual sub-

tasks.

Different from these approaches, we treat the challenge as a training data imbalance problem and advocate a pose-based data balancing strategy to address this issue.

Cascaded Networks: In the light of the coarse-to-fine cascaded regression framework, multiple networks can be stacked to form a stronger network to boost the performance. To this end, shape- or landmark-related features should be used to satisfy the training of multiple networks in cascade. However, a CNN using a global face image as input cannot meet this requirement. To address this issue, one solution is to extract CNN features from local patches around facial landmarks. This idea is advocated, for example, by Trigeorgis *et al*. who use the Recurrent Neural Network (RNN) for end-to-end model training [56]. As an alternative, we can train a network based on the global image patch for rough facial landmark localisation. Then, for each landmark or a composition of multiple landmarks in a specific region of the face, a network is trained to perform fine-grained landmark prediction [54, 14, 39, 65]. For another example, Yu *et al*. propose to inject local deformations to the estimated facial landmarks of the first network using thin-plate spline transformations [68].

In this paper, we use a two-stage CNN-based landmark localisation framework. The first CNN is a very simple one that can perform rough facial landmark localisation very quickly. The aim of the first network is to mitigate the difficulties posed by inaccurate face detection and in-plane head rotations. Then the second CNN is used to perform fine-grained landmark localisation.

3. CNN-based facial landmark localisation

The target of CNN-based facial landmark localisation is to find a nonlinear mapping:

$$\Phi : \mathcal{I} \rightarrow \mathbf{s}, \tag{1}$$

that outputs a shape vector $\mathbf{s} \in \mathbb{R}^{2L}$ for a given input colour image $\mathcal{I} \in \mathbb{R}^{H \times W \times 3}$. The input image is usually cropped using the bounding box output by a face detector. The shape vector is in the form of $\mathbf{s} = [x_1, ..., x_L, y_1, ..., y_L]^T$, where L is the number of pre-defined 2D facial landmarks and (x_l, y_l) are the coordinates of the lth landmark. To obtain this mapping, first, we have to define the architecture of a multi-layer neural network with randomly initialised parameters. In fact, the mapping $\Phi = (\phi_1 \circ ... \circ \phi_M)(\mathcal{I})$ is a composition of M functions, in which each function stands for a specific layer in the network.

Given a set of labelled training samples $\Omega = \{\mathcal{I}_i, \mathbf{s}_i\}_{i=1}^N$, the target of CNN training is to find a Φ that minimises:

$$\sum_{i=1}^{N} loss(\Phi(\mathcal{I}_i), \mathbf{s}_i), \tag{2}$$

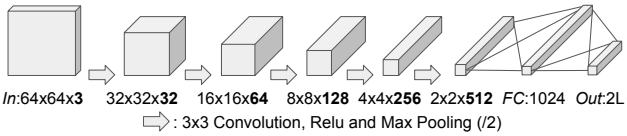


Figure 2. Our simple CNN-6 network consisting of 5 convolutional and 1 fully connected layers followed by an output layer.

where $loss()$ is a pre-defined loss function that measures the difference between a predicted shape vector and its ground truth. In such a case, the CNN is used as a regression model learned in a supervised manner. To optimise the above objective function, optimisation algorithms such as Stochastic Gradient Descent (SGD) can be used.

To empirically analyse different loss functions, we use a simple CNN architecture, in the following termed CNN-6, for facial landmark localisation, to achieve high speed in model training and testing. The input for this network is a $64 \times 64 \times 3$ colour image and the output is a vector of $2L$ real numbers for the 2D coordinates of L landmarks. As shown in Fig. 2, our CNN-6 has five 3×3 convolutional layers, a fully connected layer and an output layer. After each convolutional and fully connected layer, a standard Relu layer is used for nonlinear activation. A Max pooling after each convolutional layer is used to downsize the feature map to half of the size.

To boost the performance, more powerful network architectures can be used, such as our two-stage landmark localisation framework presented in Section 6 and the recently proposed ResNet architecture [24]. We will report the results of these advanced network architectures in Section 7. It should be highlighted that, to the best of our knowledge, this is the first time that such a deep residual network, *i.e*. ResNet-50, is used for facial landmark localisation.

4. Wing loss

The design of a proper loss function is crucial for CNN-based facial landmark localisation. However, mainly the L2 loss has been used in existing deep-neural-network-based facial landmarking systems. In this paper, to the best of our knowledge, we are the first to analyse different loss functions for CNN-based facial landmark localisation and demonstrate that the L1 and smooth L1 loss functions perform much better than the L2 loss. Motivated by our analysis, we propose a new loss function, namely Wing loss, which further improves the accuracy of CNN-based facial landmark localisation systems.

4.1. Analysis of different loss functions

Given a training image $\mathcal{I}$ and a network Φ, we can predict the facial landmarks as a vector $\mathbf{s}' = \Phi(\mathcal{I})$. The loss is

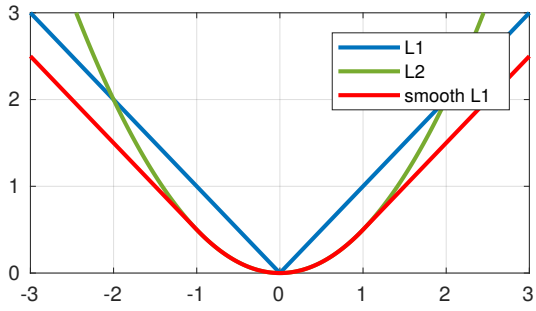


Figure 3. Plots of the L1, L2 and smooth L1 loss functions.

defined as:

$$loss(\mathbf{s}, \mathbf{s}') = \sum_{i=1}^{2L} f(s_i - s'_i), \tag{3}$$

where $\mathbf{s}$ is the ground-truth shape vector of the facial landmarks. For $f(x)$ in the above equation, L1 loss uses $L1(x) = |x|$ and L2 loss uses $L2(x) = \frac{1}{2}x^2$. The smooth L1 loss function is piecewise-defined as:

$$smooth_{L1}(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| < 1 \\ |x| - \frac{1}{2} & \text{otherwise} \end{cases}, \tag{4}$$

which is quadratic for small values of $|x|$ and linear for large values [22]. More specifically, smooth L1 uses $L2(x)$ for $x \in (-1, 1)$ and shifted $L1(x)$ elsewhere. Fig. 3 depicts the plots of these loss functions. It should be noted that the smooth L1 loss is a special case of the Huber loss [27]. The loss function that has widely been used in facial landmark localisation is the L2 loss function. However, it is well-known that the L2 loss is sensitive to outliers. This is the main reason why, *e.g.*, Girshick [22] and Rashid *et al*. [45] use the smooth L1 loss function for their localisation tasks.

For evaluation, the AFLW-Full protocol has been used [75][1]. This protocol consists of 20k training images and 4386 test images. Each image has 19 facial landmarks. We use three state-of-the-art algorithms [75, 21, 39] as our baseline for comparison. The first one is the Cascaded Compositional Learning algorithm (CCL) [75], which is a multi-view cascaded regression model based on random forests. The second one is the Two-stage Re-initialisation Deep Regression Network (TR-DRN) [39]. The last baseline algorithm is a multi-view approach based on cascaded shape regression, namely DAC-CSR [21].

We train the CNN-6 network on AFLW using three different loss functions and report the results in Table 1. The L2 loss function, which has been widely used for facial landmark localisation, performs well. The result is better than CCL in terms accuracy but worse than DAC-CSR and TR-DRN. Surprisingly, when we use L1 or smooth L1 for the CNN-6 training, the performance in terms of accuracy improves significantly and outperforms all the state-of-the-art baseline approaches, despite the CNN network's simplicity.

[1]The AFLW dataset is introduced in Section 7.2.1.

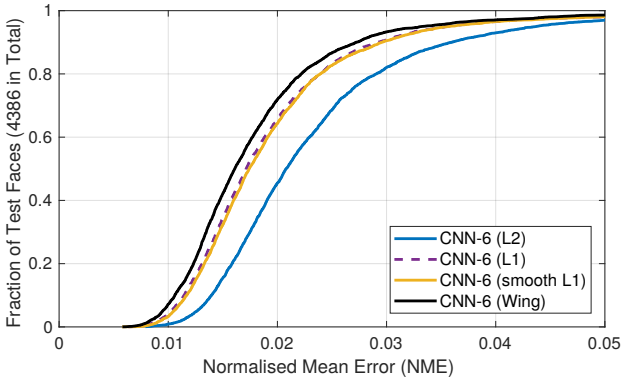


Figure 4. CED curves comparing different loss functions on the AFLW dataset, using the AFLW-Full protocol.

Table 1. A comparison of different loss functions with the three baseline algorithms in terms of the average error normalised by face size. Each training has been performed for 120k iterations. The learning rate is reduced from 3×10^{-6} to 3×10^{-8} for L2, and from 3×10^{-5} to 3×10^{-7} for the other loss functions.

method	average normalised error
CCL (CVPR2016) [75]	2.72×10^{-2}
DAC-CSR (CVPR2017) [21]	2.27×10^{-2}
TR-DRN (CVPR2017) [39]	2.17×10^{-2}
CNN-6 (L2)	2.41×10^{-2}
CNN-6 (L1)	2.00×10^{-2}
CNN-6 (smooth L1)	2.02×10^{-2}
CNN-6 (Wing loss)	1.88×10^{-2}

4.2. The proposed Wing loss

We compare the results obtained on the AFLW dataset using the simple CNN-6 network in Fig. 4 by plotting the Cumulative Error Distribution (CED) curves. We can see that all the loss functions analysed in the last section perform well for large errors. This indicates that the training of a neural network should pay more attention to the samples with small or medium range errors. To achieve this target, we propose a new loss function, namely Wing loss, for CNN-based facial landmark localisation.

In order to motivate the new loss function, we provide an intuitive analysis of the properties of the L1 and L2 loss functions (Fig. 3). The magnitude of the gradients of these two functions is 1 and $|x|$ respectively, and the magnitude of the corresponding optimal step sizes should be $|x|$ and 1. Finding the minimum in either case is straightforward. However, the situation becomes more complicated when we try to optimise simultaneously the location of multiple points, as in our problem of facial landmark localisation for-

mulated in Eq. (3). In both cases the update towards the solution will be dominated by larger errors. In the case of L1, the magnitude of the gradient is the same for all the points, but the step size is disproportionately influenced by larger errors. For L2, the step size is the same but the gradient will be dominated by large errors. Thus in both cases it is hard to correct relatively small displacements.

The influence of small errors can be enhanced by an alternative loss function, such as $\ln x$. Its gradient, given by $1/x$, increases as we approach zero error. The magnitude of the optimal step size is x^2. When compounding the contributions from multiple points, the gradient will be dominated by small errors, but the step size by larger errors. This restores the balance between the influence of errors of different sizes. However, to prevent making large update steps in a potentially wrong direction, it is important not to overcompensate the influence of small localisation errors. This can be achieved by opting for a log function with a positive offset.

This type of loss function shape is appropriate for dealing with relatively small localisation errors. However, in facial landmark detection of in-the-wild faces we may be dealing with extreme poses where initially the localisation errors can be very large. In such a regime the loss function should promote a fast recovery from these large errors. This suggests that the loss function should behave more like L1 or L2. As L2 is sensitive to outliers, we favour L1.

The above intuitive argument points to a loss function which for small errors should behave as a log function with an offset, and for larger errors as L1. Such a composite loss function can be defined as:

$$wing(x) = \begin{cases} w \ln(1 + |x|/\epsilon) & \text{if } |x| < w \\ |x| - C & \text{otherwise} \end{cases}, \quad (5)$$

where the non-negative w sets the range of the nonlinear part to $(-w, w)$, ϵ limits the curvature of the nonlinear region and $C = w - w\ln(1 + w/\epsilon)$ is a constant that smoothly links the piecewise-defined linear and nonlinear parts. Note that we should not set ϵ to a very small value because it makes the training of a network very unstable and causes the exploding gradient problem for very small errors. In fact, the nonlinear part of our Wing loss function just simply takes the curve of $\ln(x)$ between $[\epsilon/w, 1+\epsilon/w)$ and scales it along both the X-axis and Y-axis by a factor of w. Also, we apply translation along the Y-axis to allow $wing(0) = 0$ and to impose continuity on the loss function.

From Fig. 4, we can see that our Wing loss outperforms L2, L1 and smooth L1 in terms of accuracy. The Wing loss further reduces the average normalised error from 2×10^{-2} to 1.88×10^{-2}, which is 6% lower than the best result obtained in the last section (Table 1) and 13% lower than the best state-of-the-art deep-learning baseline approach, *i.e.* TR-DRN. In our experiments, we set the parameters of the Wing loss as $w = 10$ and $\epsilon = 2$. For the results of different parameter settings, please refer to Table 2.

Table 2. A comparison of different parameter settings (w and ϵ) for the proposed Wing loss function, measured in terms of the average normalised error ($\times 10^{-2}$) on AFLW using our CNN-6 network.

ϵ \ w	4	6	8	10	12	14
0.5	1.95	1.92	1.92	1.94	1.97	1.94
1	1.95	1.91	1.91	1.90	1.90	1.95
2	1.98	1.92	1.91	**1.88**	1.90	1.98
3	2.02	1.96	1.93	1.91	1.89	2.02

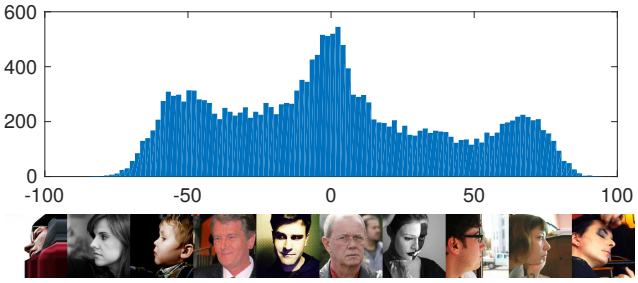


Figure 5. Distribution of the pose coefficients of the AFLW training samples by projecting their shapes to the 1-D pose space.

5. Pose-based data balancing

Extreme pose variations are very challenging for robust facial landmark localisation in the wild. To mitigate this issue, we propose a simple but very effective Pose-based Data Balancing (PDB) strategy. We argue that the difficulty for accurately localising faces with large poses is mainly due to data imbalance, which is a well-known problem in many computer vision applications [51]. For example, given a training dataset, most samples in it are likely to be near-frontal faces. The neural network trained on such a dataset is dominated by frontal faces. By over-fitting to the frontal pose it cannot adapt well to faces with large poses. In fact, the difficulty of training and testing on merely frontal faces should be similar to that on profile faces. This is the main reason why a view-based face analysis algorithm usually works well for pose-varying faces. As an evidence, even the classical view-based Active Appearance Model can localise faces with large poses very well (up to 90° in yaw) [9].

To perform PDB, we first align all the training shapes to a reference shape using Procrustes Analysis, with the mean shape as the reference shape. Then we apply PCA to the aligned training shapes and project the original shapes to the one dimensional space defined by the shape eigenvector (pose space) controlling pose variations. The distribution of projection coefficient of the training samples is represented by a histogram with K bins, plotted in Figure 5. With this histogram, we balance the training data by duplicating the

Table 3. A comparison of different loss functions using our PDB strategy and two-stage landmark localisation framework, measured in terms of the average normalised error ($\times 10^{-2}$) on AFLW. The method CNN-6/7 indicates the proposed two-stage localisation framework using CNN-6 as the first network and CNN-7 as the second network (Section 6). For CNN-7, the learning rate is reduced from 1×10^{-6} to 1×10^{-8} for L2, and from 1×10^{-5} to 1×10^{-7} for the L1, smooth L1 and Wing loss functions.

method \ loss	L2	L1	smooth L1	Wing
CNN-6	2.41	2.00	2.02	1.88
CNN-6 + PDB	2.23	1.89	1.91	1.83
CNN-6/7	2.06	1.82	1.84	1.71
CNN-6/7 + PDB	1.94	1.73	1.76	1.65

samples falling into the bins of lower occupancy. We modify each duplicated sample by performing random image rotation, bounding box perturbation and other data augmentation approaches introduced in Section 7.1. To deal with in-plane rotations, we use a two-stage facial landmark localisation framework that will be introduced in Section 6. The results obtained by the CNN-6 network with PDB are shown in Table 3. It should be noted that PDB improves the performance of CNN-6 on the AFLW dataset for all different types of loss functions.

6. Two-stage landmark localisation

Besides the out-of-plane head rotations, the accuracy of a facial landmark localisation algorithm can be degraded by other factors, such as in-plane head rotations and inaccurate bounding boxes output from a poor face detector. To mitigate this issue, we advocate the use of a two-stage landmark localisation framework.

In the proposed two-stage localisation framework, we use a very simple network, *i.e*. the CNN-6 network with $64 \times 64 \times 3$ input images, as the first network. The CNN-6 network is very fast (400 fps on an NVIDIA GeForce GTX Titan X *Pascal*), hence it will not slow down the speed of our facial landmark localisation algorithm too much. The landmarks output by the CNN-6 network are used to refine the input image for the second network by removing the in-plane head rotation and correcting the bounding box. Also, the input image resolution for the second network is increased for fine-grained landmark localisation from $64 \times 64 \times 3$ to $128 \times 128 \times 3$, with the addition of one set of convolutional, Relu and Max pooling layers. Hence, the term 'CNN-7' is used to denote the second network. The CNN-7 network has a similar architecture to the CNN-6 network in Fig. 2. The difference is that CNN-7 has 6 convolutional layers which resize the feature map from $128 \times 128 \times 3$ to $2 \times 2 \times 512$. In addition, for the first convolutional layer in CNN-7, we double the number of 3×3 kernels from 32 to 64. We use the term 'CNN-6/7' for our two-stage facial landmark localisation framework and compare it with the CNN-6 network in Table 3. As reported in the table, the use of our two-stage landmark localisation framework further improves the accuracy, regardless of the type of loss function used.

7. Experimental results

In this section, we evaluate our method on the Annotated Facial Landmarks in the Wild (AFLW) dataset [32] and the 300 Faces in the Wild (300W) dataset [49]. We first introduce our implementation details and experimental settings. Then we compare our algorithm with state-of-the-art approaches on AFLW and 300W. Last, we analyse the performance of different networks in terms of both accuracy and speed.

7.1. Implementation details

In our experiments, we used Matlab 2017a and the MatConvNet toolbox[2]. The training and testing of our networks were conducted on a server running Ubuntu 16.04 with $2\times$ Intel Xeon E5-2667 v4 CPU, 256 GB RAM and 4 NVIDIA GeForce GTX Titan X (*Pascal*) cards. Note that we only use one GPU card for measuring the run time. We set the weight decay to 5×10^{-4}, momentum to 0.9 and batch size to 8 for network training. Each model was trained for 120k iterations. We did not use any other advanced techniques in our CNN-6 and CNN-7 networks, such as batch normalisation, dropout or residual blocks. The standard ReLu function was used for nonlinear activation, and Max pooling with the stride of 2 was used to downsize feature maps. For the convolutional layer, we used 3×3 kernels with the stride of 1. All our networks, except ResNet-50, were trained from scratch without any pre-training on any other dataset. For the proposed PDB strategy, the number of bins K was set to 17 for AFLW and 9 for 300W.

For CNN-6, the input image size is $64 \times 64 \times 3$. We reduced the learning rate from 3×10^{-6} to 3×10^{-8} for the L2 loss, and from 3×10^{-5} to 3×10^{-7} for the other loss functions. The parameters of the Wing loss were set to $w = 10$ and $\epsilon = 2$. For CNN-7, the input image size is $128 \times 128 \times 3$. We reduced the learning rate from 1×10^{-6} to 1×10^{-8} for the L2 loss, and from 1×10^{-5} to 1×10^{-7} for the other loss functions. The parameters of the Wing loss were set to $w = 15$ and $\epsilon = 3$.

To perform data augmentation, we randomly rotated each training image between $[-30, 30]$ degrees for CNN-6 and between $[-10, 10]$ degrees for CNN-7. In addition, we randomly flipped each training image with the probability of 50%. For bounding box perturbation, we applied random translations to the upper-left and bottom-right corners of the face bounding box within 5% of the bounding

[2]http://www.vlfeat.org/matconvnet/

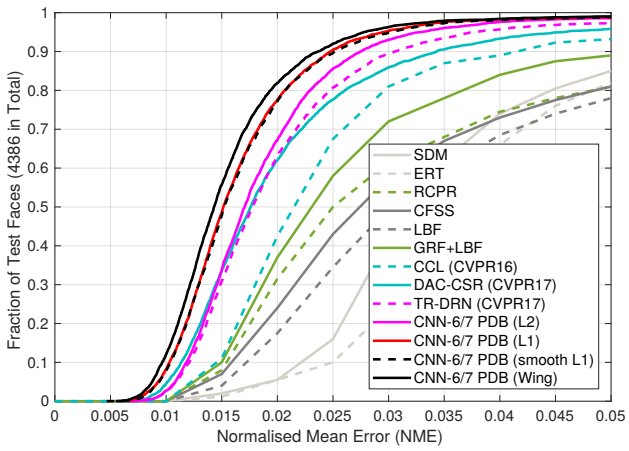


Figure 6. A comparison of the CED curves on the AFLW dataset. We compare our method with a set of state-of-the-art approaches, including SDM [63], ERT [30], RCPR [5], CFSS [74], LBF [46], GRF [23], CCL [75], DAC-CSR [21] and TR-DRN [39].

box size. Last, we randomly injected Gaussian blur ($\sigma = 1$) to each training image with the probability of 50%.

Evaluation Metric: For evaluation of a facial landmark localisation algorithm, we adopted the widely used Normalised Mean Error (NME). For the AFLW dataset using the AFLW-Full protocol, the given face bounding box of a test sample is a square [75]. To calculate the NME of a test sample, the AFLW-Full protocol uses the width (or height) of the face bounding box as the normalisation term. For the 300W dataset, we followed the protocol used in [46]. This protocol uses the inter-pupil distance as the normalisation term, which is different from the standard 300W protocol that uses the outer eye corner distance.

7.2. Comparison with state of the art

7.2.1 AFLW

We first evaluated our algorithm on the AFLW dataset [32], using the AFLW-Full protocol [75]. AFLW is a very challenging dataset that has been widely used for benchmarking facial landmark localisation algorithms. The images in AFLW consist of a wide range of pose variations in yaw (from $-90°$ to $90°$), as shown in Fig. 5. The AFLW-Full protocol contains 20,000 training and 4,386 test images, and each image has 19 manually annotated facial landmarks.

We compare the proposed method with state-of-the-art approaches in terms of accuracy in Fig. 6 using the Cumulative Error Distribution (CED) curve. In our experiments, we used our two-stage facial landmark localisation framework by stacking the CNN-6 and CNN-7 networks (denoted by CNN-6/7), as introduced in Section 6. In addition, the proposed Pose-based Data Balancing (PDB) strategy was adopted, as presented in Section 5. We report the results of the proposed approach using four different loss functions.

Table 4. A comparison of the proposed approach with the state-of-the-art approaches on the 300W dataset in terms of the NME averaged over all the test samples. We follow the protocol used in [46]. Note that the error is normalised by the inter-pupil distance, rather than the outer eye corner distance.

method \ subset	Com.	Challenge	Full
RCPR [5]	6.18	17.26	8.35
CFAN [71]	5.50	16.78	7.69
ESR [6]	5.28	17.00	7.58
SDM [63]	5.60	15.40	7.52
ERT [30]	-	-	6.40
CFSS [74]	4.73	9.98	5.76
TCDCN [73]	4.80	8.60	5.54
LBF [46]	4.95	11.98	6.32
3DDFA (CVPR16) [76]	6.15	10.59	7.01
3DDFA + SDM	5.53	9.56	6.31
DDN (ECCV16) [68]	-	-	5.65
RAR (ECCV16) [62]	*4.12*	8.35	*4.94*
DeFA (ICCVW17) [38]	5.37	9.38	6.10
TR-DRN (CVPR17) [39]	4.36	*7.56*	4.99
CNN-6/7 + PDB (L2)	4.18	8.19	4.97
CNN-6/7 + PDB (L1)	3.58	**7.02**	4.26
CNN-6/7 + PDB (smooth L1)	3.57	7.08	4.26
CNN-6/7 + PDB (Wing)	**3.27**	7.18	**4.04**

As shown in Fig. 6, our CNN-6/7 network outperforms all the other approaches even when trained with the commonly used L2 loss function (magenta solid line). This validates the effectiveness of the proposed two-stage localisation framework and the PDB strategy. Second, by simply switching the loss function from L2 to L1 or smooth L1, the performance of our method has been improved significantly (red solid and black dashed lines). Last, the use of our newly proposed Wing loss function further improves the accuracy (black solid line). The proportion of test samples (Y-axis) associated with a small to medium normalised mean error (X-axis) is increased.

7.2.2 300W

The 300W dataset is a collection of multiple face datasets, including LFPW [1], HELEN [34], AFW [77] and XM2VTS [42]. The face images involved in 300W have been semi-automatically annotated by 68 facial landmarks [50]. To perform the evaluation on 300W, we followed the protocol used in [46]. The protocol uses the full set of AFW and the training subsets of LFPW and HELEN as the training set, which contains 3148 training samples in total. The test set of the protocol includes the test subsets of LFPW and HELEN, as well as 135 IBUG face images newly collected by the managers of the 300W dataset. The final size of the test set is 689. The test set is further divided into two subsets for evaluation, *i.e.* the common and challenging subsets. The common subset has 554 face images from the LFPW and HELEN test subsets and the challenging subset constitutes the 135 IBUG face images.

Table 5. A comparison of our simple network with ResNet-50, in terms of accuracy on AFLW-Full and 300W.

	AFLW	300W		
		Com.	Challenge	Full
CNN-6 + PDB (Wing)	1.83	3.35	7.20	4.10
CNN-6/7 + PDB (Wing)	1.65	3.27	7.18	4.04
ResNet-50 + PDB (Wing)	1.47	3.01	6.01	3.60

Table 6. A comparison in accuracy of ResNet-50 using different loss functions, evaluated on AFLW-Full.

Loss Function	L2	L1	smooth L1	Wing
NME ($\times 10^{-2}$)	1.68	1.51	1.52	1.47

Similar to the experiments conducted on the AFLW dataset, we used the two-stage localisation framework with our PDB strategy. The results obtained by our approach with different loss functions are reported in Table 4.

As shown in Table 4, our two-stage landmark localisation framework with the PDB strategy and the newly proposed Wing loss function outperforms all the other state-of-the-art algorithms on the 300W dataset in accuracy. The error has been reduced by almost 20% as compared to the current best result reported by the RAR algorithm [62].

7.3. Run time and network architectures

Facial landmark localisation has been widely used in many real-time practical applications, hence the speed together with accuracy of an algorithm is crucial for the deployment of the algorithm in commercial use cases.

To analyse the performance of our Wing loss on more advanced network architectures, we evaluated ResNet [24] for the task of landmark localisation on AFLW and 300W. We used the ResNet-50 model that was pre-trained on the ImageNet ILSVRC classification problem[3]. We fine-tuned the model on the training sets of AFLW and 300W separately for landmark localisation. The input for ResNet is a $224 \times 224 \times 3$ colour image. It should be highlighted that, to our best knowledge, this is the first time that such a deep network has been used for facial landmark localisation.

For both AFLW and 300W, by replacing the CNN-6/7 network with ResNet-50, the performance has been further improved by around 10%, as shown in Table 5. However, this performance boosting comes at the cost of much slower training and inference of ResNet compared to CNN-6/7.

To validate the effectiveness of our Wing loss for large capacity networks, we also conducted experiments using ResNet-50 with different loss functions on AFLW. The results are reported in Table. 6. The results further demonstrate the superiority of the proposed Wing loss over other loss functions for large capacity networks, *e.g*. ResNet-50.

Last, we evaluated the speed of different networks on the 300W dataset with 68 landmarks for both GPU and CPU devices. The results are reported in Table 7. According to the table, our simple CNN-6/7 network is roughly an order of magnitude faster than ResNet-50 at the compromise of 10% performance difference in accuracy. Also, our CNN-6/7 model is much faster than most existing DNN-based facial landmark localisation approaches such as TR-DRN [39]. The speed of TR-DRN is 83 fps on an NVIDIA GeForce GTX Titan X card. Even with a powerful GPU card, it is hard to achieve video rate (60fps) with ResNet-50. It should be noted that our CNN-6/7 still outperforms the state-of-the-art approaches by a significant margin while running at 170 fps on a GPU card, as shown in Fig. 6.

[3]http://www.vlfeat.org/matconvnet/pretrained/

Table 7. A comparison of different networks, in the number of model parameters, model size and speed.

network	# params	size	speed (fps)	
			GPU	CPU
CNN-6	3.8 M	14 MB	400	150
CNN-6/7	12.3 M	46 MB	170	20
ResNet-50	25 M	99 MB	30	8

8. Conclusion

In this paper, we analysed different loss functions that can be used for the task of regression-based facial landmark localisation. We found that L1 and smooth L1 loss functions perform much better in accuracy than the L2 loss function. Motivated by our analysis of these loss functions, we proposed a new, Wing loss performance measure. The key idea of the Wing loss criterion is to increase the contribution of the samples with small and medium size errors to the training of the regression network. To prove the effectiveness of the proposed Wing loss function, extensive experiments have been conducted using several CNN network architectures. Furthermore, a pose-based data balancing strategy and a two-stage landmark localisation framework were advocated to improve the accuracy of CNN-based facial landmark localisation further. By evaluating our algorithm on multiple well-known benchmarking datasets, we demonstrated the merits of the proposed approach.

It should be emphasised that the proposed Wing loss is relevant to other regression-based computer vision tasks using convolutional neural networks. However, being constrained by the space limitations, we leave the discussion of its extended use to future reports.

Acknowledgements

This work was supported in part by the EPSRC Programme Grant (FACER2VM) EP/N007743/1, EPSRC/dstl/MURI project EP/R018456/1, the National Natural Science Foundation of China (61373055, 61672265) and the NVIDIA GPU Grant Program.

References

[1] 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 (CVPR)*, pages 545–552, 2011.

[2] C. Bhagavatula, C. Zhu, K. Luu, and M. Savvides. Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[3] A. Bulat and G. Tzimiropoulos. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[4] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[5] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1513–1520, 2013.

[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. Cootes, C. Taylor, D. Cooper, J. Graham, et al. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[8] T. F. Cootes, G. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[9] T. F. Cootes, K. Walker, and C. J. Taylor. View-based active appearance models. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 227–232, 2000.

[10] 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.

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

[12] J. Deng, G. Trigeorgis, Y. Zhou, and S. Zafeiriou. Joint multi-view face alignment in the wild. *arXiv preprint arXiv:1708.06023*, 2017.

[13] 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.

[14] Y. Dong and Y. Wu. Adaptive cascade deep convolutional neural networks for face alignment. *Computer Standards & Interfaces*, 42:105–112, 2015.

[15] P. Dou, S. K. Shah, and I. A. Kakadiaris. End-to-end 3d face reconstruction with deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[16] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[17] 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.

[18] 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.

[19] Z.-H. Feng, P. Huber, J. Kittler, P. Hancock, X.-J. Wu, Q. Zhao, P. Koppen, and M. Rätsch. Evaluation of Dense 3D Reconstruction from 2D Face Images in the Wild. *arXiv preprint arXiv:1803.05536*, 2018.

[20] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu. Face detection, bounding box aggregation and pose estimation for robust facial landmark localisation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 160–169, 2017.

[21] Z.-H. Feng, J. Kittler, W. Christmas, P. Huber, and X.-J. Wu. Dynamic Attention-Controlled Cascaded Shape Regression Exploiting Training Data Augmentation and Fuzzy-Set Sample Weighting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2481–2490, 2017.

[22] R. Girshick. Fast R-CNN. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1440–1448, 2015.

[23] 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.

[24] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[25] 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.

[26] P. Huber, P. Kopp, W. Christmas, M. Rätsch, and J. Kittler. Real-time 3d face fitting and texture fusion on in-the-wild videos. *IEEE Signal Processing Letters*, 24(4):437–441, 2017.

[27] P. J. Huber et al. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[28] A. Jourabloo and X. Liu. Large-Pose Face Alignment via CNN-Based Dense 3D Model Fitting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[29] A. Jourabloo, M. Ye, X. Liu, and L. Ren. Pose-invariant face alignment with a single cnn. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[30] 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.

[31] J. Kittler, P. Huber, Z.-H. Feng, G. Hu, and W. Christmas. 3D Morphable Face Models and Their Applications. In *Inter-*

national Conference on Articulated Motion and Deformable Objects, pages 185–206. Springer, 2016.

[32] 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.

[33] P. Koppen, Z.-H. Feng, J. Kittler, M. Awais, W. Christmas, X.-J. Wu, and H.-F. Yin. Gaussian Mixture 3D Morphable Face Model. *Pattern Recognition*, 74:617–628, 2018.

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

[35] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[36] Z. Liang, S. Ding, and L. Lin. Unconstrained facial landmark localization with backbone-branches fully-convolutional networks. *arXiv preprint arXiv:1507.03409*, 2015.

[37] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[38] Y. Liu, A. Jourabloo, W. Ren, and X. Liu. Dense face alignment. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017.

[39] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[40] I. Masi, S. Rawls, G. Medioni, and P. Natarajan. Pose-aware face recognition in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[41] I. Masi, A. T. Tran, T. Hassner, J. T. Leksut, and G. Medioni. Do we really need to collect millions of faces for effective face recognition? In *European Conference on Computer Vision (ECCV)*, pages 579–596, 2016.

[42] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The extended M2VTS database. In *International Conference on Audio and Video-based Biometric Person Authentication*, volume 964, pages 965–966. Citeseer, 1999.

[43] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 483–499, 2016.

[44] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[45] M. Rashid, X. Gu, and Y. Jae Lee. Interspecies knowledge transfer for facial keypoint detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[46] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment via regressing local binary features. *IEEE Transactions on Image Processing*, 25(3):1233–1245, 2016.

[47] S. Romdhani, S. Gong, A. Psarrou, et al. A multi-view nonlinear active shape model using kernel PCA. In *British Machine Vision Conference (BMVC)*, volume 99, pages 483–492, 1999.

[48] J. Roth, Y. Tong, and X. Liu. Adaptive 3d face reconstruction from unconstrained photo collections. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[49] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 Faces in-the-Wild Challenge: The first facial landmark localization Challenge. In *International Conference on Computer Vision - Workshops (ICCVW)*, pages 397–403, 2013.

[50] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 896–903, 2013.

[51] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 761–769, 2016.

[52] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun. Cascaded hand pose regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 824–832, 2015.

[53] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3476–3483, 2013.

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

[55] 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.

[56] 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.

[57] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic. Copula ordinal regression for joint estimation of facial action unit intensity. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[58] Y. Wu, C. Gou, and Q. Ji. Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[59] Y. Wu, T. Hassner, K. Kim, G. Medioni, and P. Natarajan. Facial landmark detection with tweaked convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[60] Y. Wu and Q. Ji. Constrained joint cascade regression framework for simultaneous facial action unit recognition and fa-

cial landmark detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[61] Y. Wu, S. K. Shah, and I. A. Kakadiaris. GoDP: Globally Optimized Dual Pathway deep network architecture for facial landmark localization in-the-wild. *Image and Vision Computing*, 2017.

[62] 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)*, pages 57–72, 2016.

[63] 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.

[64] X. Xiong and F. De la Torre. Global supervised descent method. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2664–2673, 2015.

[65] X. Xu and I. A. Kakadiaris. Joint head pose estimation and face alignment framework using global and local cnn features. In *IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pages 642–649, 2017.

[66] J. Yang, Q. Liu, and K. Zhang. Stacked hourglass network for robust facial landmark localisation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2025–2033, 2017.

[67] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua. Neural aggregation network for video face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[68] X. Yu, F. Zhou, and M. Chandraker. Deep deformation network for object landmark localization. In *European Conference on Computer Vision (ECCV)*, pages 52–70, 2016.

[69] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[70] 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)*, 2016.

[71] 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 (ECCV)*, volume 8690, pages 1–16, 2014.

[72] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

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

[74] 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.

[75] 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)*, pages 3409–3417, 2016.

[76] 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)*, pages 146–155, 2016.

[77] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2879–2886, 2012.