Not all points are created equal - an anisotropic cost function for facial landmark location

Farshid Rayhan¹ f.rayhan@manchester.ac.uk Aphrodite Galata¹ A.Galata@manchester.ac.uk Tim F Cootes² timothy.f.cootes@manchester.ac.uk ¹ Dept of Computer Science The University of Manchester Manchester, UK

² School of Health Sciences The University of Manchester Manchester, UK

Abstract

An effective approach to locating facial landmarks is to train a CNN to predict their positions directly from an image patch cropped around the face. Earlier work has shown that the choice of cost function comparing predicted with target points is important, but have tended to use the same weighting for each individual point. Since some points, such as those on boundaries, are less clearly defined than those at obvious corners, we propose an alternative cost function which uses anisotropic weights. This penalises movement away from feature boundaries more than that along them. We demonstrate that using this cost function improves location performance and training convergence. We also address the problem of pose imbalance in datasets, suggesting a way of balancing the poses in the training samples. State of the art results on three public datasets (AFLW, WFLW and 300W) demonstrate the effectiveness of these techniques

1 Introduction

Facial landmark localisation ("face alignment") aims to find the coordinates of a set of predefined key points for 2D face images. Each landmark usually has specific semantic meaning, such as an eye corner or nose tip, which provides rich geometric information for tasks such as 3D face reconstruction [19, 23, 25, 26, 27, 50, 52, 47], face recognition [53, 51, 53], and emotion estimation [13, 54, 54].

Traditional approaches to locating points include Active Shape Models (ASM) [5], Active Appearance Models (AAM) [5], Constrained Local Models (CLM) [2] and cascaded-regression-based approaches [11, 13, 14, 14, 14, 15, 15, 15]. Recently, deep convolutional neural networks have been used [14, 13, 14, 14, 15], 15], 15].

To perform robust face alignment using deep neural networks, different types of network have been explored including Convolutional Neural Networks (CNN) [50], Recurrent Neural Networks (RNN) [52, 52] and Auto-Encoder Networks [59]. Various network architectures have been extensively studied, for example, Fully Convolutional Networks (FCN) [53] and hourglass networks with residual blocks have been found very effective [9, 51].

Deep neural networks can be trained to predict the point positions directly from an image patch of the whole face. Recent work has shown that this approach can give good results, but that the choice of the cost function is important [21, 59].

In this work we propose a simple, but effective, anisotropic cost function which leads to improved performance, in part by downweighting the contribution of points which are less well defined in the training set. We also proposed a method of dealing with the imbalance in poses in training sets by resampling. We evaluate the performance of these techniques on three widely used datasets, and demonstrate that combining them achieves state-of-the-art performance.

2 Related Work

There are two broad approaches to facial landmark location using CNNs;

- To search for each point by scanning over a region with a classifier or local regressor
 producing probability maps for each point,
- 2. to predict the point positions directly with regression on the whole face patch [20, 59].

This work addresses the second approach.

Network Architectures: CNN-based facial landmark localisation approaches are mostly regression-based. For such a task, the most straightforward way is to use a CNN with fully connected final layers [13, 13, 51], where the input is usually a image of a face and the output is a vector consisting the 2D coordinates of the landmarks. In recent years, researchers have proposed CNN systems addressing this problem and shown promising results using Hourglass networks [1, 1, 2, 11, 56] and FCN [53]. They are different from traditional CNNs because they output a pseudo-probability maps for each facial landmarks. In [51], a distance aware softmax function was proposed to reduce the false alarms in such generated 2D maps.

Pose Variations: Variation in head pose can make the location of facial points more challenging. Many strategies have been proposed to address this issue. Multiview models have been used to improve the performance of many approaches. In [12], a multiview cascaded regression model is trained using a fuzzy membership weighting strategy that performs better than some CNN based approaches. Another popular strategy is the use of 3D face models [2, 23, 23, 56, 72]. By estimating the pose of a input 2D face image and recovering the 3D shape the problem can be addressed. 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 [16, 59, 73]. Multi-task learning has been adopted to address the difficulties posed by pose variations [12, 54, 75].

Cascaded Networks: It has been found that stacking multiple networks to form a stronger network can boost the performance. To this end, landmark or shape related features are used to address the training of multiple networks in a cascade. One approach for faces is for each network in the cascade to attempt to improve the position of a landmark point based on the image information around the current position. For instance, [\Box] used a Recurrent Neural Network (RNN) for end-to-end training. Alternatively, a network can be trained based on the global image patch for an initial landmark localisation. Then, for each landmark or a group of landmarks in a specific region of the face, a network is trained to perform fine-grained landmark prediction [\Box], \Box , \Box , \Box]. In [\Box], the authors have proposed system

where they inject local deformations to the estimated facial landmarks of the first network using thin-plate spline transformations.

Loss functions: L1 and L2 loss functions have been used widely to train different types of model for facial landmark localisation [5, 49, 50]. For deep-neural-network-based facial landmarking systems [1] L2 losses have been preferred. In [1], the authors provide an analysis where they conclude that L1 and a smoothed L1 loss performs much better that L2. In [22] a novel loss function ("Wing-Loss") is proposed, that performs significantly better than traditional L1, L2 and smooth L1 loss and was able to achieve the lowest error on recent test datasets.

In this paper, we use a one-stage end-to-end trainable CNN-based facial landmark localisation framework. We construct a cost function which takes account of the likely uncertainty in the training point positioning. Rather than measuring errors using Euclidean distances and treating all points equally, we use the Mahalanobis distance with anisotropic covariance matrices, so each point may be given a different weight. The covariance matrices for points on edges are constructed so as to penalise movement away from the boundary more than that along the boundary. Landmarks such as those at the corners of the eyes or mouth are usually well defined in both x and y directions, so are given unit covariance matrices. This relatively simple modification leads to both better overall accuracy, and to faster convergence during training.

3 **Methods**

3.1 **Cost Function**

Suppose we have two shapes each containing *n* points, $\{\mathbf{x}_i\}$ and $\{\mathbf{z}_i\}$. If we assume isotropic weights, equal for all points (i.e. we penalise all displacements equally), then a cost function comparing them has the form

$$Q_1 = \sum_{i=1}^{n} |\mathbf{x}_i - \mathbf{z}_i|^2 \tag{1}$$

Suppose now that we assume a co-variance matrix, S_i at each point $\{z_i\}$ which indicates the expected spread of errors at each point.

Then a natural cost function is

$$Q_2 = \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{z}_i)^T \mathbf{W}_i (\mathbf{x}_i - \mathbf{z}_i)$$
(2)

where $\mathbf{W}_i = \mathbf{S}_i^{-1}$, defining a weight matrix at each point.

We can construct a covariance matrix to represent anisotropic distributions, with a variance of a^2 along direction **u** (where $|\mathbf{u}| = 1$) and b^2 along the orthogonal direction. In this case.

$$\mathbf{S} = a^2 \mathbf{u} \mathbf{u}^T + b^2 (\mathbf{I} - \mathbf{u} \mathbf{u}^T), \text{ so } \mathbf{W} = \mathbf{S}^{-1} = \frac{1}{a^2} \mathbf{u} \mathbf{u}^T + \frac{1}{b^2} (\mathbf{I} - \mathbf{u} \mathbf{u}^T)$$
(3)

If we are matching points on two curves, we typically require a stronger constraint (thus lower variance) along the direction normal to the curve, compared to that along the curve we are less concerned about points sliding along the curves.

Assume that the points $\{z_i\}$ are roughly equally spaced along a curve. Then for all points but the end points, the tangent at point *i* is given approximately by

$$\mathbf{t}_{i} = (\mathbf{z}_{i+1} - \mathbf{z}_{i-1}) / |\mathbf{z}_{i+1} - \mathbf{z}_{i-1}|$$
(4)

At the ends of an open curve we use

$$\mathbf{t}_1 = (\mathbf{z}_2 - \mathbf{z}_1) / |\mathbf{z}_2 - \mathbf{z}_1|, \text{ and } \mathbf{t}_n = (\mathbf{z}_n - \mathbf{z}_{n-1}) / |\mathbf{z}_n - \mathbf{z}_{n-1}|$$
 (5)

We set up the weight matrices as

$$\mathbf{W}_i = \mathbf{S}_i^{-1} = \frac{1}{a^2} \mathbf{t}_i \mathbf{t}_i^T + \frac{1}{b^2} (\mathbf{I} - \mathbf{t}_i \mathbf{t}_i^T)$$
(6)

where a > b, giving more freedom to slide along the curve than normal to it. See Figure 1. At well defined corners we use $\mathbf{W}_i = \mathbf{I}$.

We find that this cost function leads to more rapid convergence. In Figure 2, we display outputs of a ResNet50 after the first epoch on the same subject. It shows that using our loss, the model has already learned facial shape within a single epoch whereas the model trained with Wing loss [21] gives poor point predictions.

3.2 Data balancing

Dealing with wide pose variation is a challenging issue for facial landmark localisation. In [20], is was shown that balancing the dataset based on pose improved the prediction quality significantly. Where most of the images in a training dataset are frontal faces, the network over-fits to the frontal shape which leads to poor performance on faces far from frontal.

[\square] used Procrustes Analysis to align all the shapes to the reference shape (ie. mean shape). Then they used PCA to project the original shapes into a low dimensional space, which was divided into K pose bins. Samples in bins with with lower occupancy are replicated to improve the balance of different poses.

We followed [22]'s data balancing strategy and divide the training set into *K* bins based on pose. We calculated the mean shape for each pose bin, $\bar{\mathbf{x}}_k$, then the overall mean of these individual means: $\bar{\mathbf{x}} = \frac{1}{K} \sum_{k=1}^{K} \bar{\mathbf{x}}_k$.

When training, we then predict the displacement from this overall mean, $\delta \mathbf{x} = \mathbf{x} - \bar{\mathbf{x}}$. This helps speed up convergence without affecting overall convergence.

During training we *undersample* the training set using [\square] where we find the bin with the lowest number of samples and randomly choosing same number of samples from other bins. We used the same number of bins, K = 9 for 300W, as recommended in [\square].

4 **Experiments**

4.1 Datasets

We tested our approach on the AFLW [5], 300W [3], 300W private test dataset and the WFLW [5] dataset. The WFLW dataset is the most challenging. We follow the same protocol as [13, 19, 20, 5] on those datasets for our experiments. For 300W and AFLW, we also report the prediction score of [20, 5] to demonstrate that we have created the same environment. Details about the dataset and implementation details can be found in supplementary materials.

4.2 Evaluation Metrics

Normalised Mean Error (NME) is a commonly used [12, 19, 20] metric to evaluate the quality of facial alignment algorithms. The NME is defined as:

$$NME(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{M} \sum_{i}^{M} \frac{||\mathbf{x}_{i} - \hat{\mathbf{x}}_{i}||}{d}$$
(7)

where **x** and $\hat{\mathbf{x}}$ are the ground truth and the predicted landmarks respectively, *M* is the number of landmarks of each image, $\hat{\mathbf{x}}$ is the *i*th predicted landmark, and *d* is a normalisation factor. For the AFLW and WFLW dataset, we set *d* as the inter-pupil (distance of eye centers). For the 300W dataset, we provide results for both inter-ocular distance (distance of outer eye corners) used as the original evaluation protocol in [12], and inter-pupil distance used in [12]. For the WFLW dataset, we use the inter-ocular distance described in [13].

Failure Rate (FR) is a metric to evaluate localisation quality. For one image, if NME is larger than a threshold, *t*, then it is considered a failed prediction. For the 300W testset, we use t = 8% and t = 10% to compare with different approaches. We follow [20, 51] and use 10% as the threshold for the WFLW dataset.

Cumulative Error Distribution (CED) curve shows the proportion of examples with an NME below t, as a function of t. The curve is plotted from zero up to the NME failure rate threshold (e.g. 10%, 8%) and the Area Under Curve (AUC) is calculated based on the CED curve. A higher AUC indicates that larger portion of the test set has been well predicted.



Figure 1: Error margin for each point where the red ellipsoids represent the freedom of movement along the curve

4.3 Evaluation on AFLW

We first evaluated our algorithm on the AFLW dataset, using the AFLW-Full protocol [22]. AFLW is a challenging dataset which has been widely used for benchmarking other facial landmark localisation systems [19, 20, 52, 53].

For AFLW dataset, we followed the protocol from 'adaptive wing loss' [5] to generate all 14 boundary lines. This also make the performance of Anisotropic loss comparable. We compare Anisotropic loss with state-of-the-art approaches in terms of NME(%) in table 2. Our method is able to outperform the state-of-the-art methods which proves its robustness to large pose variations.

4.4 Evaluation on 300W

Anisotropic loss is able to achieve the state-of-the-art performance on the 300W testing

RAYHAN: NOT ALL POINTS ARE CREATED EQUAL

Method		Common	Challenging	Fullset		
			Subset			
Inter-pupil Normalisation						
DVLN-CVPR17	[57]	3.94	7.62	4.66		
TSR-CVPR17	[37]	4.36	7.56	4.99		
DSRN-CVPR18 [4.12	9.68	5.21		
DCFE-ECCV18 [3.83	7.54	4.55		
LAB-CVPR18	[56]	3.42	6.98	4.12		
Wing-CVPR18 [20]		3.27	7.18	4.04		
Wing-CVPR18* [20]		3.31	7.20	4.17		
Awing-CVPR 19 [53		3.77	6.52	4.31		
Anisotropic loss *		$\textbf{3.12} \pm \textbf{0.2}$	$\textbf{6.25} \pm \textbf{0.47}$	$\textbf{3.94} \pm \textbf{0.34}$		
Inter-ocular Normalisation						
PCD-CNN-CVPR 18	[28]	3.67	7.62	4.44		
CPM+SBR-CVPR 18	[[]]]	3.28	7.58	4.1		
SAN-CVPR 18		3.34	6.6	3.98		
LAB-CVPR 18	[56]	2.98	5.19	3.49		
Awing-CVPR 19	[53]	2.72	4.52	3.07		
Awing-CVPR 19*	[53]	2.8	4.58	3.12		
Anisotropic loss *		$\textbf{2.35} \pm \textbf{0.15}$	$\textbf{4.05} \pm \textbf{0.5}$	$\textbf{2.91} \pm \textbf{0.22}$		

Table 1: NME results on the 300W testset. We report mean and standard deviation of Anisotropic loss for 5 runs. '*' indicates the experiments we ran ourselves, showing that the results in published works were reproducible.

Method	Full(%)	Frontal(%)		
CCLCVPR 16 [2.72	2.17		
TSR CVPR 17 [57]	2.17	-		
DAC-OSR CVPR 17 [🗳]	2.27	1.81		
DCFE ECCV 18 [53]	2.17	-		
CPM+SBR CVPR 18 [2.14 -			
SAN CVPR 18 [🖽]	1.91	1.85		
DSRN CVPR 18 [1.86	-		
LAB CVPR 18 [56]	1.85	1.62		
Wing CVPR 18 [20]	1.65	-		
RCN+(L+ELT+A) CVPR 18 [1.59	-		
AWing [53]	1.53	1.38		
Anisotropic loss	$\textbf{1.3} \pm \textbf{0.09}$	$\textbf{1.27} \pm \textbf{0.1}$		
Table 2: Normalised Mean error(%) on the AFLW testset				

dataset using NME metric, see Table 1. For the challenging subset (iBug dataset), we are able to outperform Adaptive Wing loss [53] and wing loss [21] by a significant margin. Furthermore, on the 300W private test dataset (see Table 5), Anisotropic loss performs better than the previous state-of-the-art on various metrics including NME, AUC and FR measured at both 8% NME and 10% NME (See also the detailed table for 300W and 300W private dataset in supplementary materials). Note that we almost halved the failure rate of the next best baseline to 0.26%.

4.5 Evaluation on WFLW

Our method achieves the best results on the WFLW dataset, as shown in Table 3. This set is significantly more difficult than AFLW and 300W. We outperformed the previous state-of-the-art on FR and AUC on every subset by a significant margin.

RAYHAN: NOT ALL POINTS ARE CREATED EQUAL

Metric	Method	Testset	Pose Subset	Expression Subset	Illumination Subset	Make-up Subset	Occlusion Subset	Blur Subset
	SDM CVPR 13 [10.29	24.10	11.45	9.32	9.38	13.03	11.28
	CFSS CVPR 15 [9.07	21.36	10.09	8.30	8.74	11.76	9.96
NME(%)	DVLN CVPR 17 [6.08	11.54	6.78	5.73	5.98	7.33	6.88
	LAB CVPR 18 [11]	5.27	10.24	5.51	5.23	5.15	6.79	6.32
	Wing CVPR 18 [5.11	8.75	5.36	4.93	5.41	6.37	5.81
	AWing CVPR 19 [53]	4.36	7.38	4.58	4.32	4.27	5.19	4.96
	Anisotropic loss (GTBbox)	4.01	6.87	4.17	4.15	4	4.94	3.84
	SDM CVPR 13 [29.40	84.36	33.44	26.22	27.67	41.85	35.32
	CFSS CVPR 15 [20.56	66.26	23.25	17.34	21.84	32.88	23.67
FR10% (%)	DVLNCVPR 17 [10.84	46.93	11.15	7.31	11.65	16.30	13.71
	LAB CVPR 18 [11]	7.56	28.83	6.37	6.73	7.77	13.72	10.74
	Wing CVPR 18 [22]	6.00	22.70	4.78	4.30	7.77	12.50	7.76
	AWing CVPR 19 [2.84	13.50	2.23	2.58	2.91	5.98	3.75
	Anisotropic loss (GTBbox)	2.34	13.11	2.04	2.24	2.35	5.67	3.52
-	SDM CVPR 13 [0.3002	0.0226	0.2293	0.3237	0.3125	0.2060	0.2398
	CFSS CVPR 15 [0.3659	0.0632	0.3157	0.3854	0.3691	0.2688	0.3037
AUC10%	DVLNCVPR 17 [0.4551	0.1474	0.3889	0.4743	0.4494	0.3794	0.3973
	LAB CVPR 18 [55]	0.5323	0.2345	0.4951	0.5433	0.5394	0.4490	0.4630
	Wing CVPR 18 [22]	0.5504	0.3100	0.4959	0.5408	0.5582	0.4885	0.4918
	AWing CVPR 19 [0.5719	0.3120	0.5149	0.5777	0.5715	0.5022	0.5120
	Anisotropic loss (GTBbox)	0.5895	0.3247	0.5252	0.5901	0.5786	0.5344	0.5297

Table 3: Evaluation on the WFLW dataset. GTBbox indicates the ground truth landmarks are used to crop faces.

Method	Common	Challenging	Fullset
	Subset	Subset	
Inter-pupil 1	Normalisation		
Awing-CVPR 19 [53]	3.77	6.52	4.31
Anisotropic loss	$3.68\pm0.08*$	6.44 ± 0.19	$4.21 \pm 0.09*$
Anisotropic loss + Direct prediction	$3.64\pm0.07*$	6.42 ± 0.18	$4.20 \pm 0.07*$
Anisotropic loss + data balancing from [22]	$3.52 \pm 0.21*$	$6.32\pm0.53^*$	$4.12 \pm 0.37^{*}$
Anisotropic loss + our data balancing	3.12 ± 0.2	$\textbf{6.25} \pm \textbf{0.47}$	$\textbf{3.94} \pm \textbf{0.34}$

Table 4: Ablation study on the 300W testset using different data balancing method using Anisotropic loss (5 runs). '*' indicates results superior to state of the art result and **bold** values represents the best result

5 Analysis/ablation study

Figure 2 shows results on an image from the 300W dataset with a network (a ResNet50) trained for 0-3 epochs. Here we observe that on epoch 0 (with no training) the CNN trained with both loss Wing Loss and Anisotropic loss predicts all the landmarks which are on the top left corner. It demonstrates that our cost function seems to converge to good solutions more quickly than the Wingloss cost. By epoch 4, using our loss function, the network has already found the rough location of each landmarks where as the network trained using Wingloss is still struggling. In another example, in figure 3, we can see that after the first epoch the 2D points have a front facing face-shape even though the input face is facing a side, however within 30 epochs the predicted points are much more accurate. Following this pattern, the CNN converges after 10k epochs when trained with Anisotropic loss .

Anisotropic loss allows some freedom of movement (see figure: 1) along the curve but constrains movement orthogonal to the curve. We set a variance of a^2 along the curve and b^2 orthogonal to it. We used the 300W dataset to examine the effect of varying these two parameters. We found that the best results are obtained for a = 5, b = 3 (see full details in supplementary materials). Note that corner points are given a unit covariance matrix, so are effectively given significantly more weight in the cost function than points on curves.

In Table 6 we show the performance of similar methods (a more detailed comparison is given in the supplementary material). We compare Anisotropic loss with other state of the art loss functions on 300W, AFLW and WFLW datasets. We report values from WingLoss [22] and Adaptive WingLoss [53]. We observe that our method performs better that all other



Figure 2: A comparison between Wingloss (top row) and our loss (bottom row) where the outputs from a ResNet50 after being trained for 0 to 3 epochs. The yellow dots are the ground truths and blue dots are predictions.

Method	NME	AUC8%	FR8%
DAN CVPRW 17 [4.30	47.00	2.67
SHN CVPRW17 [🖬]	4.05	-	-
DCFE ECCV 18 [53]	3.88	52.42	1.83
AWing CVPR 19 [53]	3.56	55.76	0.83
AWing CVPR 19* [53]	3.6	55.7	0.9
Anisotropic loss *	$\textbf{3.15}\pm\textbf{0.2}$	$\textbf{56.87} \pm \textbf{0.13}$	$\textbf{0.49} \pm \textbf{0.18}$
	NME	AUC10%	FR10%
DR + MDM CVPR 17 [-	52.19	3.67
JMFA17Õ [🛛]	-	54.85	1.00
LAB CVPR 18 [55]	-	58.85	0.83
AWing CVPR 19 [53]	3.56	64.40	0.33
AWing CVPR 19* [🗳]	3.6	64.45	0.4
Anisotropic loss *	3.15 ± 0.2	$\textbf{66.08} \pm \textbf{0.09}$	0.26 ± 0.11

Table 5: Evaluation on the 300W private dataset. We report mean and standard deviation of Anisotropic loss for 5 runs. '*' indicates the experiments we ran.

5 loss functions.

We tried a range of network architectures, including ResNet [2], DenseNet [2], Wide ResNet [2] on 300W, AFLW and WFLW dataset using Anisotropic loss . We found that there was little difference between them (Full details are in the supplementary material). We used the ResNet50 for all our other experiments.

We compared the performance of Anisotropic loss with and without various data balancing techniques (see Table 4). Anisotropic loss without any data balancing was able achieve state of the art NME score in the common and full set of the 300W dataset. Unsurprisingly there is almost no difference between predicting the points directly compared to predicting the displacement from the mean. Using sampling technique from [20] improved the performance by quite a margin but also increased the standard deviation of errors. Anisotropic loss

dataset	L2	L1	Smooth L1	Wing loss	Adaptive Wing loss	Anisotropic loss
300W	5.12	4.98	4.58	4.04*	4.17*	$\textbf{3.94} \pm \textbf{0.34}$
AFLW	1.94*	1.73*	1.76*	1.65*	1.53*	$\textbf{1.3} \pm \textbf{0.09}$
WFLW	10.12	9.45	6.08	5.11*	4.36*	4.01
	· .	0. 41.00				

Table 6: A comparison of different loss functions (L2, L1, smooth L1, Wing loss, Adaptive Wingloss and Anisotropic loss) using of NME (inner pupil norm) on 300W and AFLW dataset using ResNet50. '*' indicates reported values









Epoch 0Epoch 10Epoch 20Epoch 30Figure 3: Example on a non-front facing face using Anisotropic loss . The yellow dots are
the ground truths and blue dots are predictions.

performed the best with our data balancing technique. While the variation is larger when our data balancing is used compared to no sampling, it is less than that found using pose balancing from [22].

6 Limitations

The proposed cost function requires a method of estimating the shape of the covariance for each point. We have assumed the same values for a, b for every point, but it might be better to have individual values for each point - these could be estimated from the amount of curvature of any boundary at the point (the higher the curvature, the lower the variance).

7 Conclusion

We have introduced a novel loss function, Anisotropic loss, for regression based facial landmark localisation and a pose balancing scheme. We compute a separate covariance matrix for each point using the formula in equation 4. By "curves" we mean piece-wise lines through the points around boundaries of structures (chin, mouth, nose etc). We will change one of the figures to make the construction clearer. The covariance matrices are pre-computed and remain fixed throughout the training. Though these are relatively simple modification, they lead to a significant improvement in overall performance on three widely used datasets.

8 Acknowledgement

This project is funded by Toyota Motor Europe and The University of Manchester. We would also like to thank Gabriel Othmezouri (gabriel.othmezouri@toyota-europe.com) and Marleen De Weser (marleen.de.weser@toyota-europe.com) for all their inputs

References

 Riza Alp Guler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou, and Iasonas Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6799–6808, 2017.

- [2] Chandrasekhar Bhagavatula, Chenchen Zhu, Khoa Luu, and Marios Savvides. Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3980–3989, 2017.
- [3] Adrian Bulat and Georgios Tzimiropoulos. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3706–3714, 2017.
- [4] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017.
- [5] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.
- [6] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681– 685, 2001.
- [7] David Cristinacce and Timothy F Cootes. Feature detection and tracking with constrained local models. In *Bmvc*, volume 1, page 3. Citeseer, 2006.
- [8] Adrian K Davison, Claudia Lindner, Daniel C Perry, Weisang Luo, Timothy F Cootes, et al. Landmark localisation in radiographs using weighted heatmap displacement voting. In *International Workshop on Computational Methods and Clinical Applications* in *Musculoskeletal Imaging*, pages 73–85. Springer, 2018.
- [9] Jiankang Deng, George Trigeorgis, Yuxiang Zhou, and Stefanos Zafeiriou. Joint multiview face alignment in the wild. *IEEE Transactions on Image Processing*, 28(7):3636– 3648, 2019.
- [10] Piotr Dollár, Peter Welinder, and Pietro Perona. Cascaded pose regression. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 1078–1085. IEEE, 2010.
- [11] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *Proceedings of the IEEE Conference on Computer Vision* and Pattern Recognition, pages 379–388, 2018.
- [12] Xuanyi Dong, Shoou-I Yu, Xinshuo Weng, Shih-En Wei, Yi Yang, and Yaser Sheikh. Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In *Proceedings of the IEEE Conference on Computer Vision* and Pattern Recognition, pages 360–368, 2018.
- [13] Yuan Dong and Yue Wu. Adaptive cascade deep convolutional neural networks for face alignment. *Computer standards & interfaces*, 42:105–112, 2015.

- [14] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5562–5570, 2016.
- [15] Zhen-Hua Feng, Patrik Huber, Josef Kittler, William Christmas, and Xiao-Jun Wu. Random cascaded-regression copse for robust facial landmark detection. *IEEE Signal Processing Letters*, 22(1):76–80, 2014.
- [16] Zhen-Hua Feng, Guosheng Hu, Josef Kittler, William Christmas, and Xiao-Jun 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.
- [17] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Face detection, bounding box aggregation and pose estimation for robust facial landmark localisation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 160–169, 2017.
- [18] Zhen-Hua Feng, Josef Kittler, William Christmas, Patrik Huber, and Xiao-Jun Wu. Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2481–2490, 2017.
- [19] Zhen-Hua Feng, Patrik Huber, Josef Kittler, Peter Hancock, Xiao-Jun Wu, Qijun Zhao, Paul Koppen, and Matthias Rätsch. Evaluation of dense 3d reconstruction from 2d face images in the wild. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pages 780–786. IEEE, 2018.
- [20] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2235–2245, 2018.
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [22] Sina Honari, Pavlo Molchanov, Stephen Tyree, Pascal Vincent, Christopher Pal, and Jan Kautz. Improving landmark localization with semi-supervised learning. In *Proceedings* of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1546– 1555, 2018.
- [23] Guosheng Hu, Fei Yan, Josef Kittler, William Christmas, Chi Ho Chan, Zhenhua Feng, and Patrik Huber. Efficient 3d morphable face model fitting. *Pattern Recognition*, 67: 366–379, 2017.
- [24] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

- [25] Patrik Huber, Philipp Kopp, William Christmas, Matthias Rätsch, and Josef Kittler. Real-time 3d face fitting and texture fusion on in-the-wild videos. *IEEE Signal Processing Letters*, 24(4):437–441, 2016.
- [26] Lei Jiang, Xiao-Jun Wu, and Josef Kittler. Dual attention mobdensenet (damdnet) for robust 3d face alignment. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.
- [27] Lei Jiang, Xiao-Jun Wu, and Josef Kittler. Robust 3d face alignment with efficient fully convolutional neural networks. In *International Conference on Image and Graphics*, pages 266–277. Springer, 2019.
- [28] Amin Jourabloo and Xiaoming Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4188–4196, 2016.
- [29] Amin Jourabloo, Mao Ye, Xiaoming Liu, and Liu Ren. Pose-invariant face alignment with a single cnn. In *Proceedings of the IEEE International Conference on computer vision*, pages 3200–3209, 2017.
- [30] Josef Kittler, Patrik Huber, Zhen-Hua Feng, Guosheng Hu, and William Christmas. 3d morphable face models and their applications. In *International Conference on Articulated Motion and Deformable Objects*, pages 185–206. Springer, 2016.
- [31] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In 2011 IEEE international conference on computer vision workshops (ICCV workshops), pages 2144–2151. IEEE, 2011.
- [32] Paul Koppen, Zhen-Hua Feng, Josef Kittler, Muhammad Awais, William Christmas, Xiao-Jun Wu, and He-Feng Yin. Gaussian mixture 3d morphable face model. *Pattern Recognition*, 74:617–628, 2018.
- [33] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 88–97, 2017.
- [34] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep localitypreserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017.
- [35] Zhujin Liang, Shengyong Ding, and Liang Lin. Unconstrained facial landmark localization with backbone-branches fully-convolutional networks. *arXiv preprint arXiv:1507.03409*, 2015.
- [36] Yaojie Liu, Amin Jourabloo, William Ren, and Xiaoming Liu. Dense face alignment. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 1619–1628, 2017.

- [37] Jiangjing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, and Xi Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3317–3326, 2017.
- [38] Iacopo Masi, Stephen Rawls, Gérard Medioni, and Prem Natarajan. Pose-aware face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4838–4846, 2016.
- [39] Iacopo Masi, Anh Tun Trn, Tal Hassner, Jatuporn Toy Leksut, and Gérard Medioni. Do we really need to collect millions of faces for effective face recognition? In *European Conference on Computer Vision*, pages 579–596. Springer, 2016.
- [40] Xin Miao, Xiantong Zhen, Xianglong Liu, Cheng Deng, Vassilis Athitsos, and Heng Huang. Direct shape regression networks for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5040– 5049, 2018.
- [41] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.
- [42] Rajeev Ranjan, Vishal M Patel, and Rama 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*, 41 (1):121–135, 2017.
- [43] Maheen Rashid, Xiuye Gu, and Yong Jae Lee. Interspecies knowledge transfer for facial keypoint detection. In *Proceedings of the IEEE Conference on Computer Vision* and Pattern Recognition, pages 6894–6903, 2017.
- [44] Farshid Rayhan, Sajid Ahmed, Asif Mahbub, Rafsan Jani, Swakkhar Shatabda, and Dewan Md Farid. Cusboost: cluster-based under-sampling with boosting for imbalanced classification. In 2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), pages 1–5. IEEE, 2017.
- [45] Farshid Rayhan, Aphrodite Galata, and Timothy F Cootes. Choicenet: Cnn learning through choice of multiple feature map representations. *arXiv preprint arXiv:1904.09472*, 2019.
- [46] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014.
- [47] Joseph Roth, Yiying Tong, and Xiaoming Liu. Adaptive 3d face reconstruction from unconstrained photo collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4197–4206, 2016.
- [48] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013.

- [49] Xiao Sun, Yichen Wei, Shuang Liang, Xiaoou Tang, and Jian Sun. Cascaded hand pose regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 824–832, 2015.
- [50] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013.
- [51] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE* conference on computer vision and pattern recognition, pages 1701–1708, 2014.
- [52] George Trigeorgis, Patrick Snape, Mihalis A Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic descent method: A recurrent process applied for endto-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision* and Pattern Recognition, pages 4177–4187, 2016.
- [53] Roberto Valle, Jose M Buenaposada, Antonio Valdes, and Luis Baumela. A deeplyinitialized coarse-to-fine ensemble of regression trees for face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 585–601, 2018.
- [54] Robert Walecki, Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. Copula ordinal regression for joint estimation of facial action unit intensity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4902–4910, 2016.
- [55] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6971–6981, 2019.
- [56] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE* conference on computer vision and pattern recognition, pages 2129–2138, 2018.
- [57] Wenyan Wu and Shuo Yang. Leveraging intra and inter-dataset variations for robust face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 150–159, 2017.
- [58] Yue Wu and Qiang Ji. Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3400–3408, 2016.
- [59] Yue Wu, Chao Gou, and Qiang Ji. Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion. In *Proceedings of the IEEE conference* on computer vision and pattern recognition, pages 3471–3480, 2017.
- [60] Yue Wu, Tal Hassner, KangGeon Kim, Gerard Medioni, and Prem Natarajan. Facial landmark detection with tweaked convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):3067–3074, 2017.
- [61] Yuhang Wu, Shishir K Shah, and Ioannis A Kakadiaris. Godp: Globally optimized dual pathway deep network architecture for facial landmark localization in-the-wild. *Image and Vision Computing*, 73:1–16, 2018.

- [62] Shengtao Xiao, Jiashi Feng, Junliang Xing, Hanjiang Lai, Shuicheng Yan, and Ashraf Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *European conference on computer vision*, pages 57–72. Springer, 2016.
- [63] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.
- [64] Xiang Xu and Ioannis A Kakadiaris. Joint head pose estimation and face alignment framework using global and local cnn features. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pages 642–649. IEEE, 2017.
- [65] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. In *Proceedings* of the IEEE conference on computer vision and pattern recognition, pages 4362–4371, 2017.
- [66] Jing Yang, Qingshan Liu, and Kaihua Zhang. Stacked hourglass network for robust facial landmark localisation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 79–87, 2017.
- [67] Xiang Yu, Feng Zhou, and Manmohan Chandraker. Deep deformation network for object landmark localization. In *European Conference on Computer Vision*, pages 52– 70. Springer, 2016.
- [68] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.
- [69] Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Occlusion-free face alignment: Deep regression networks coupled with de-corrupt autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3428–3437, 2016.
- [70] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.
- [71] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4998–5006, 2015.
- [72] Shizhan Zhu, Cheng Li, Chen-Change Loy, and Xiaoou Tang. Unconstrained face alignment via cascaded compositional learning. In *Proceedings of the IEEE Conference* on Computer Vision and Pattern Recognition, pages 3409–3417, 2016.
- [73] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer* vision and pattern recognition, pages 146–155, 2016.