Mining Hard Augmented Samples for Robust Facial Landmark Localisation with CNNs

Zhen-Hua Feng, *Member, IEEE,* Josef Kittler, *Life Member, IEEE,* and Xiao-Jun Wu, *Member, IEEE,*

***Abstract*—Effective data augmentation is crucial for facial landmark localisation with Convolutional Neural Networks (CNNs). In this letter, we investigate different data augmentation techniques that can be used to generate sufficient data for training CNN-based facial landmark localisation systems. To the best of our knowledge, this is the first study that provides a systematic analysis of different data augmentation techniques in the area. In addition, an online Hard Augmented Example Mining (HAEM) strategy is advocated for further performance boosting. We examine the effectiveness of those techniques using a regression-based CNN architecture. The experimental results obtained on the AFLW and COFW datasets demonstrate the importance of data augmentation and the effectiveness of HAEM. The performance achieved using these techniques is superior to the state-of-the-art algorithms.**

***Index Terms*—Facial landmark localisation, deep neural networks, data augmentation, hard augmented example mining.**

I. Introduction

Facial landmark localisation is one of the key preprocessing steps in facial image analysis. Typical facial landmark localisation methods include active shape model [1], active appearance model [2], and cascaded shape regression [3], [4]. These methods have been extensively studied during the past decades [5]–[10]. More recently, deep neural networks have become the mainstream in the area [11]–[16], especially in unconstrained scenarios characterised by the presence of a wide spectrum of appearance variations, *e.g.* in pose, expression, illumination and occlusion. One prerequisite to successful training of a network is a huge amount of labelled data. However, to label a large dataset with tens of landmarks per face manually is difficult and tedious. To avoid this problem, training data augmentation has become an essential alternative.

The aim of data augmentation is to increase the diversity of an existing training set, and thus to improve the generalisation capability of a trained network to unseen samples. Typical data augmentation approaches for facial landmark localisation inject geometric and textural variations by techniques such as image flip, rotation, scale, translation, blurring and colour jetting, applied to an input image. These augmentation approaches are very efficient to implement thus can be easily performed online for network training. However, to the best of our knowledge, the impact of these data augmentation methods on the performance of a facial landmark localisation system has not been properly investigated. To fill this gap, we introduce different data augmentation approaches in this paper and perform a systematic analysis of their effectiveness in the context of a regression-based Convolutional Neural Network (CNN). Based on our preliminary experiments, we show that not all augmentation techniques are equally effective in contributing to the training set diversity. Cubuk *et. al.* reported similar findings and proposed 'AutoAugment' that is able to find the best augmentation policy for an arbitrary training dataset by searching a large number of randomly generated policies [17]. In this paper, to mine the most useful augmented samples for the training of a CNN on an arbitrary dataset, we advocate an online Hard Augmented Example Mining (HAEM) strategy for performance boosting.

It should be noted that the spectrum of data augmentation approaches used in facial landmark localisation includes computationally expensive techniques such as synthesising a large number of virtual face images with head rotations based on a 3D face model [18]. These have the capacity to inject further degrees of diversity to an existing training dataset. However, such data augmentation is normally performed offline to maximise the network training speed. We could also perform efficient online data augmentation strategies offline, but they would require a lot of storage because we have to save the augmented training samples for each mini-batch. In this paper, we focus on simple methods that are efficient for online training data augmentation. For more studies of complex data augmentation approaches that are commonly performed offline, the reader is referred to [18]–[20]. The main contributions of this paper include:

- A comprehensive study of different data augmentation approaches that can be used for facial landmark localisation. The findings provide useful guidelines for future studies in CNN-based facial landmark localisation.
- A novel online HAEM strategy for selecting effective augmented samples for network training.
- A deep analysis of different data augmentation techniques as well as the proposed HAEM strategy. The experimental results on the AFLW and COFW datasets demonstrate the merits of the proposed method.

The rest of this paper is organised as follows: In Section II, we introduce the methodology of regression-based facial landmark localisation with CNNs. We discuss different data augmentation approaches in Section III and present the

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, 61876072, 61602390) and the NVIDIA GPU Grant Program.

Z.-H. Feng and J. Kittler are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK. (email: {z.feng, j.kittler}@surrey.ac.uk.)

X.-J. Wu is with the Jiangsu Provincial Laboratory of Pattern Recognition and Computational Intelligence, Jiangnan University, Wuxi, China. (email: wu_xiaojun@jiangnan.edu.cn.)

(a) input (b) Gaussian blur (c) salt and pepper noise (d) colour jetting (e) occlusion (f) flip (g) bounding box perturbation (h) rotation (i) shear

Fig. 1. Examples of different data augmentation approaches. Each augmentation method is applied to the input image with a random parameter.

HAEM method in Section IV. The experimental results are reported in Section V and conclusions are drawn in Section VI.

II. Facial Landmark Localisation with CNNs

Given a face image, we first crop the face region using the bounding box output from a face detector. Then the cropped face region is resized to a unified size as the input of a CNN. The resized colour image is mathematically represented as a 3rd order tensor, $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$, where H and W are the image height and width. The task of landmark localisation with regression CNNs is to find a nonlinear mapping:

$$\Phi(\phi_1 \circ ... \circ \phi_M) : \mathcal{I} \rightarrow \mathbf{s}, \tag{1}$$

where Φ is a deep neural network with M layers, $\mathbf{s} = [x_1, ..., x_L, y_1, ..., y_L]^T$ is a face shape vector consisting of L pre-defined facial landmarks and (x_l, y_l) are the coordinates of the lth facial landmark.

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

$$\frac{1}{2L}\sum_{i=1}^{N}\sum_{j=1}^{L}|x_{i,j}^* - x_{i,j}'| + |y_{i,j}^* - y_{i,j}'|, \tag{2}$$

where * and $'$ denote the ground truth value and estimated value of the CNN for the corresponding facial landmark coordinate. Here we use the L1 loss as it provides better performance in accuracy than the widely used L2 loss [16]. To optimise the above objective function, in this paper, we use the Stochastic Gradient Descent (SGD) method.

III. Training Data Augmentation

In this section, we first introduce a number of data augmentation approaches. Then we empirically validate the impact of these approaches on the accuracy of a CNN-based facial landmark localisation system. We divide the data augmentation techniques into two categories: textural and geometric augmentation. Typical textural data augmentation approaches include Gaussian blur, salt and pepper noise, colour jetting and random occlusion. Geometric data augmentation approaches consist of horizontal image flip, bounding box perturbation, rotation and shear transformation. Some examples of different data augmentation methods are shown in Fig. 1.

A. Textural Data Augmentation

1) Gaussian Blur: We apply Gaussian blur to an image to smooth the image and reduce the noise as well as texture details. A positive parameter, σ, is used to control the extent of Gaussian blur. This parameter stands for the standard deviation of a Gaussian blurring function. In practice, given a predefined value of σ, we randomly sample a value between $[0, \sigma]$ using the uniform distribution to perform Gaussian blur.

2) Salt and Pepper Noise: We add salt and pepper noise to an image to increase the robustness of a trained CNN. A parameter, $\tau \in [0, 1]$, is used to define the percentage of the pixels replaced by a random intensity value. Similar to Gaussian blur, given a predefined controlling parameter τ, we use the uniform distribution to select a value between $[0, \tau]$ to apply salt and pepper noise.

3) Colour Jetting: To perform colour jetting, we adjust the intensity values of each RGB channel by saturating a predefined percentage of top and bottom pixel values and map the others to $[0, 255]$. Given a predefined parameter, $\eta \in [0, 0.5)$, we select a random high-pass threshold η_{min} between $[0, \eta]$ for each colour channel and set the values of the pixels that have the values smaller than $255 \times \eta_{min}$ to 0. We also randomly select a low-pass threshold η_{max} between $[1 - \eta, 1]$ and set the pixel values greater than $255 \times \eta_{max}$ to 255. The greylevel values of all the remaining pixels are remapped from $[\eta_{min} \times 255, \eta_{max} \times 255]$ to $[0, 255]$.

4) Occlusion: To inject occlusion, we first randomly select a pixel in the image as the centre. Then an image patch with random pixel values is used to replace the region around the centre. The width/height of the image patch is randomly set to a value between $[0, w/h \times \gamma]$ using the uniform distribution, where w is the input image width, h is the height and $\gamma \in [0, 1]$ is a predefined control parameter.

B. Geometric Data Augmentation

1) Image Flip: Flipping an image is very straightforward and has been widely used in facial landmark localisation. With such a simple operation, we can easily double the size of a given training dataset. It should be noted that the order of the landmarks is changed after the image flipping from left to right. For example, the order of the outer corners of the left and right eyes are switched after flipping an image. All the other textural or geometric image augmentation methods do not change the order of facial landmarks.

2) Bounding Box Perturbation: To perform bounding box perturbation, we randomly shift the left-upper and right-bottom corners of a bounding box along the X-axis and Y-axis. The shifted value is set to a random number between $[-w \times \alpha, w \times \alpha]$ for X-axis and $[-h \times \alpha, h \times \alpha]$ for Y-axis, where w and h are the width and height of the original bounding box, $\alpha \in [0, 0.5)$ is a pre-defined parameter. We perform image padding for

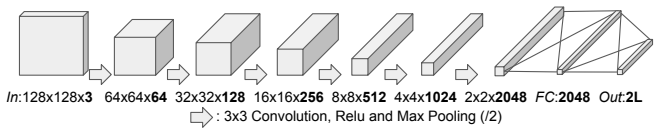

Fig. 2. The details of our CNN architecture with 6 convolutional layers, 1 fully connected layer and 1 output layer.

TABLE I
A COMPARISON OF DIFFERENT **TEXTURAL** AUGMENTATION METHODS ON AFLW-FULL, MEASURED IN NME ($\times 10^{-2}$).

Method	No Aug.	Gau. blur	S&P Noise	Col. Jet.	Occlu.
Setting	-	$\sigma = 1$	$\tau = 0.2$	$\eta = 0.4$	$\gamma = 0.6$
NME	2.093	2.077	2.090	2.009	1.990

TABLE II
A COMPARISON OF DIFFERENT **GEOMETRIC** AUGMENTATION METHODS ON AFLW-FULL, MEASURED IN NME ($\times 10^{-2}$).

Method	No Aug.	Flip	Bbox Pertu.	Rotation	Shear
Setting	-	-	$\alpha = 0.15$	$\theta = 25$	$\beta = 0.3$
NME	2.093	1.986	1.720	1.757	1.821

TABLE III
A COMPARISON OF THE USE OF COMPOUNDED AUGMENTATION METHODS ON AFLW-FULL, MEASURED IN NME ($\times 10^{-2}$).

Method	No Aug.	All Tex.	All Geo.	All Tex. & Geo.
NME	2.093	1.978	1.641	1.622

pixels at the boundary when a perturbed bounding box is out of the image.

3) Image Rotation: For an input image, we rotate the image as well as the landmarks by randomly sampling a value between $[-\theta, \theta]$, where $\theta \in [0, 180]$ is a pre-defined controlling parameter of image rotation.

4) Shear Transform: We define a matrix:

$$\begin{bmatrix} 1 & c_x & 0 \\ c_y & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

to perform a shear transformation, where the values of c_x and c_y are randomly sampled between $[-\beta, \beta]$ with uniform distribution, and $\beta \in [0, 1]$ is a predefined parameter.

C. Analysis of Different Augmentation Approaches

To examine the impact of different data augmentation approaches on the performance of a facial landmark localisation system with regression-based CNNs, we use the AFLW-Full dataset that has 20,000 training images and 4386 test images with 19 manually annotated facial landmarks per image. More details of the AFLW-Full dataset are given in Section V. We design a simple CNN architecture for the experiments. The input for the network is a $128 \times 128 \times 3$ colour image and the output is a shape vector with $2L$ real numbers. As shown in Fig. 2, the network has six 3×3 convolutional layers, one fully connected layer and one output layer. After each convolutional and fully connected layer, the nonlinear Relu function is applied. In addition, Max pooling is used after each convolutional layer to down-sample the resolution of a feature map to its half size. We use the Mean Error normalised by face size (NME) as our evaluation metric [21].

We first test each augmentation method separately on the AFLW-Full dataset. The localisation errors of our CNN architecture with different textural or geometric data augmentation techniques, as well as the augmentation process control parameters, are shown in Table I and Table II. Note that all the augmentation parameters have been set to their best values. According to our results, almost all the data augmentation approaches improve the accuracy of our CNN model as compared with the baseline method without data augmentation. In addition, one important finding is that geometric data augmentation is more effective than textural data augmentation for the training of a regression-based CNN model.

In practice, we usually use different data augmentation approaches for training CNN models jointly. We report the performance of our CNN architecture in Table III by applying all the textural, all the geometric, and all the textural plus geometric data augmentation techniques simultaneously. We can see that, by compounding all the data augmentation approaches, the localisation error of the CNN-based facial landmark localisation system can further be reduced.

Algorithm 1 Hard Augmented Example Mining (HAEM)

repeat
  Shuffle all the training samples and split them into M mini-batches $B_{i=1}^{M}$;
  Set the hard sample set H to empty;
  for $m = 1, ..., M$ **do**
    Apply the compounded augmentation with all the textural and geometric augmentation types to each sample in the mth mini-batch, B_m, with the probability of 50% and use the augmented sample to replace the original one, resulting in a new set $\hat{B}_m$ that has the same size of B_m;
    Use the set $\{\hat{B}_m, H\}$ to perform network update and return the loss incurred by each sample in $\hat{B}_m$;
    Update H by selecting the samples in $\hat{B}_m$ contributing the 2nd to the $K+1$st largest losses;
  end for
until reach the maximum epoch iterations.

IV. HARD AUGMENTED EXAMPLE MINING

With the training data augmentation, we are able to generate a huge number of additional samples for training a CNN facial landmark localisation system. Note that these samples are generated by applying random textural and geometric variations to the original labelled training images. Some augmented samples may be harder and more effective for the training of a deep neural network and some may be less effective. To select the most effective augmented training samples, we propose an online Hard Augmented Example Mining (HAEM) strategy. The proposed HAEM strategy is summarised in Algorithm 1. In essence, we select as K hard samples from the current mini-batch those which exhibit the largest losses, but excluding the

TABLE IV
A COMPARISON WITH THE STATE-OF-THE-ART METHODS ON **AFLW**, MEASURED IN NME.

Method \ Setting	NME ($\times 10^{-2}$)	
	AFLW-full	AFLW-frontal
RCPR [23]	3.73	2.87
ERT [24]	4.35	2.75
LBF [25]	4.25	2.74
CFSS [26]	3.92	2.68
CCL (CVPR16) [21]	2.72	2.17
DAC-CSR (CVPR17) [27]	2.27	1.81
TR-DRN (CVPR17) [28]	2.17	-
Zeng *et. al.* (TIP18) [29]	2.60	-
CPM+SBR (CVPR18) [30]	2.14	-
SAN (CVPR18) [31]	1.91	1.85
GoDP (IVC18) [32]	1.84	-
our CNN	2.09	1.71
our CNN + Aug.	1.62	1.34
our CNN + Aug. + HAEM	**1.58**	**1.31**

TABLE V
A COMPARISON WITH THE STATE-OF-THE-ART METHODS ON **COFW** IN TERMS OF NME AND FAILURE RATE.

Method \ Metric	NME ($\times 10^{-2}$)	Failure Rate (%)
ESR [34]	11.2	36
RCPR [23]	8.50	20
HPM [35]	7.50	13
RCRC [6]	7.30	12
CCR [18]	7.03	10
RAR (ECCV16) [36]	6.03	4.14
Wu *et. al.* (CVPR17) [37]	6.40	-
DAC-CSR (CVPR17) [27]	6.03	4.73
Zeng *et. al.* (TIP18) [29]	8.10	19
HOSRD (TPAMI18) [38]	6.80	13
RSR (TPAMI18) [39]	5.63	-
our CNN	7.79	14.4
our CNN + Aug.	**4.88**	2.37
our CNN + Aug. + HAEM	**4.88**	**2.17**

one of dominant loss. The main reason for this conservative approach is that some of the samples generated by our random data augmentation may be too hard. Such samples become 'outliers' that disturb the convergence of the training process. Thus in each mini-batch we identify the $K+1$ hardest samples and discard the hardest one to define the hard sample set H.

V. EXPERIMENTAL RESULTS

In this section, we compare our CNN architecture with state-of-the-art facial landmark localisation algorithms on two datasets: AFLW [22] and COFW [23]. We first present our implementation details and then report the evaluation results.

A. Implementation Details

To train our CNN network, we use the SGD optimisation method. We set the weight decay to 5×10^{-4}, momentum to 0.9 and batch size to 8. We linearly reduce the learning rate from 1×10^{-3} to 1×10^{-5} for 500,000 iterations. For our HAEM strategy, we set $K = 5$ as the number of selected hard samples in each mini-batch. For each training sample in a mini-batch, except the hard samples copied from the last mini-batch, we apply all the data augmentation methods with the probability of 50%. The network was implemented by Matlab 2018a with the MatConvNet toolbox. The training of our network was conducted on a machine 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.

B. Evaluation on AFLW and COFW

We first evaluated our algorithm on the AFLW dataset [22], following the protocol used in [21]. The protocol defines 20,000 training and 4,386 test images, and each image has 19 landmarks. The evaluation is performed using two different settings: AFLW-Full and AFLW-Frontal. AFLW-Full evaluates an algorithm using all the test images, whereas AFLW-Frontal evaluates an algorithm using only frontal faces.

The COFW [23] has 1345 training images and 507 test images. Each COFW face image has 29 manually annotated facial landmarks. The COFW dataset is an extended benchmarking dataset created from the LFPW dataset [33] by adding more challenging facial images with heavy occlusions.

We compare our method with state-of-the-art algorithms using the Normalised Mean Error (NME) metric on AFLW and COFW. The error was normalised by face size and inter-ocular distance for AFLW and COFW, respectively. In addition, we also measure the failure rate on COFW using the ratio of test samples with the errors higher than 10% NME. The results are shown in Table IV and Table V. We can see that, with data augmentation, the accuracy of our CNN architecture significantly improves and beats the state-of-the-art algorithms both on AFLW and COFW. In addition, by using our HAEM strategy, the accuracy of our CNN model has further been improved on AFLW. However, for COFW, the accuracy of the network is saturated and the use of HAEM only reduces the failure rate. The experimental results obtained on AFLW and COFW validate the importance of various training data augmentation approaches as well as the proposed hard augmented sample selection strategy.

VI. CONCLUSION

In this letter, we investigated a number of data augmentation approaches that can be used to train regression CNNs for facial landmark localisation. To the best of our knowledge, this is the first study that systematically analyses different data augmentation methods. We found experimentally that geometric data augmentation methods perform much better than textural data augmentation. This provides guidelines for selecting data augmentation approaches in future studies. In addition, to select more effective samples, we proposed an online augmented sample mining strategy. The experimental results indicate that the proposed method is effective for CNN-based facial landmark localisation, as evidenced by its superior localisation accuracy compared to the state-of-the-art algorithms.

REFERENCES

[1] T. Cootes, C. Taylor, D. Cooper, J. Graham *et al.*, "Active shape models-their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[2] T. F. Cootes, G. Edwards, and C. J. Taylor, "Active appearance models," in *European Conference on Computer Vision*, vol. 1407, 1998, pp. 484–498.

[3] P. Dollár, P. Welinder, and P. Perona, "Cascaded pose regression," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 1078–1085.

[4] 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)*, 2013, pp. 532–539.

[5] I. Matthews and S. Baker, "Active Appearance Models Revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[6] 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*, vol. 22, no. 1, pp. 76–80, Jan 2015.

[7] H. Yang, X. Jia, I. Patras, and K.-P. Chan, "Random Subspace Supervised Descent Method for Regression Problems in Computer Vision," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1816–1820, 2015.

[8] H.-S. Lee and D. Kim, "Tensor-Based Active Appearance Model," *IEEE Signal Processing Letters*, vol. 15, pp. 565–568, 2008.

[9] 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)*, 2017, pp. 160–169.

[10] Z.-H. Feng and J. Kittler, "Advances in facial landmark detection," *Biometric Technology Today*, vol. 2018, no. 3, pp. 8–11, 2018.

[11] 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)*, 2013, pp. 3476–3483.

[12] 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)*, vol. 8690, 2014, pp. 1–16.

[13] 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.

[14] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[15] S. Zhang, H. Yang, and Z.-P. Yin, "Transferred deep convolutional neural network features for extensive facial landmark localization," *IEEE Signal Processing Letters*, vol. 23, no. 4, pp. 478–482, 2016.

[16] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, USA*, 2018, pp. 2235–2245.

[17] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *arXiv preprint arXiv:1805.09501*, 2018.

[18] 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*, vol. 24, no. 11, pp. 3425–3440, 2015.

[19] 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)*, 2016, pp. 146–155.

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

[21] 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)*, 2016, pp. 3409–3417.

[22] 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.

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

[24] 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)*, 2014, pp. 1867–1874.

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

[26] 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)*, 2015, pp. 4998–5006.

[27] 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)*, 2017, pp. 2481–2490.

[28] 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.

[29] J. Zeng, S. Liu, X. Li, D. A. Mahdi, F. Wu, and G. Wang, "Deep context-sensitive facial landmark detection with tree-structured modeling," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2096–2107, 2018.

[30] X. Dong, S.-I. Yu, X. Weng, S.-E. Wei, Y. Yang, and Y. Sheikh, "Supervision-by-Registration: An unsupervised approach to improve the precision of facial landmark detectors," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[31] X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Style Aggregated Network for Facial Landmark Detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[32] 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.

[33] 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)*, 2011, pp. 545–552.

[34] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by Explicit Shape Regression," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2887–2894.

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

[36] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim, "Robust facial landmark detection via recurrent attentive-refinement networks," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 57–72.

[37] Y. Wu, C. Gou, and Q. Ji, "Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3471–3480.

[38] J. Xing, Z. Niu, J. Huang, W. Hu, X. Zhou, and S. Yan, "Towards robust and accurate multi-view and partially-occluded face alignment," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 987–1001, 2018.

[39] Z. Cui, S. Xiao, Z. Niu, S. Yan, and W. Zheng, "Recurrent shape regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018.