A Robust Facial Landmark Detector with Mixed Loss

Xian Zhang[1,2], Xinjie Tong[1], Ziyu Li[1,2], and Wankou Yang[1,2](✉)

[1] School of Automation, Southeast University, Nanjing 210096, China
wkyang@seu.edu.cn

[2] Key Lab of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Southeast University, Nanjing 210096, China

Abstract. Facial landmark detection is one of the most important tasks in face image and video analysis. Existing algorithms based on deep convolutional neural networks have achieved good performance in public benchmarks and practical applications such as face verification, expression analysis, beauty applications and so on. However, the performance of a facial landmark detector degrades significantly when dealing with challenging facial images in the presence of extreme appearance variations such as pose, expression, occlusion, etc. To mitigate these difficulties, we propose a robust facial landmark detection algorithm based on coordinates regression in an end-to-end training fashion. By using the soft-argmax function, the network weights can be optimised with a mixed loss function. The online pose-based data augmentation technology is used to effectively solve the data imbalance problem and improve the robustness of the proposed method. Experiments conducted on the 300-W and AFLW datasets demonstrate that the performance of the proposed algorithm is competitive to the state-of-the-art heatmap regression algorithms, in terms of accuracy. Besides, our method achieves real-time speed on 300-W with 68 landmarks, which runs at 85 FPS on a Tesla v100 GPU.

Keywords: Facial landmark detection · Mixed loss · Soft-argmax · Pose-based data augmentation

1 Introduction

Facial landmark detection, also known as face alignment [29,43], is a fundamental task in various facial image and video analysis applications [31,32,41,44–46]. During the past decades, the facial landmark detection area has made significant progress. Nevertheless, many existing approaches have difficulties in dealing with in-the-wild faces with extreme appearance variations in pose, expression, illumination, blur and occlusion.

Existing facial landmark detection algorithms can be roughly divided into three categories: global appearance based approaches, constrained local models and regression-based methods. Global appearance based methods detect the key

Z. Cui et al. (Eds.): IScIDE 2019, LNCS 11935, pp. 249–261, 2019.
https://doi.org/10.1007/978-3-030-36189-1_21

points using the whole facial textural information and global shape information [3–5,13,25,30]. Constrained local model [17] is based on global face shape and independent local textural information around each key point that captures more robust information for illumination and occlusion variations. Regression-based methods can be divided into direct regression, cascade regression and regression with deep neural networks. At present, the most widely used and the most accurate methods are all based on deep Convolutional Neural Networks (CNNs) [12,16]. In this paper, the proposed facial landmark detection method is based on CNNs as well.

The key innovations of the proposed method include:

- For data augmentation, we adopt the online Pose-based Data Balancing (PDB) [8] method that balances the original training dataset. To be more specific, we copy the samples of low proportion defined by PDB and randomly modify the samples (flip, rotate, blur, etc.) including changing the copied samples with different styles since the intrinsic variance of image styles can also affects the performance of a trained network [6].
- The baseline of this paper is CPM [34] that generates a heatmap image as the final output of a network. In order to apply Wing Loss that is specially designed for coordinates regression models in this work, we introduce the soft-argmax function [21]. The function converts heatmaps to coordinates thus the network is differentiable.
- The original Wing Loss function [8] focuses on small and medium errors, but pays less attention to the samples with large errors. To address this issue, we design a new loss function, namely mixed loss, that considers the samples with errors at various magnitudes.

2 Related Work

2.1 Pose Variation

The aim of data augmentations is to reduce the bias in network training due to the imbalance of a training dataset. STN [22] applies spatial transformer network to learn transformation parameters thus to automatically initialise a training dataset. SAN [6] translates each image to four different styles by a generative adversarial module. Both of them try to inject diversity to a training dataset and balance the training samples.

2.2 Regression Model

The regression methods used from facial landmark detection can be divided into two categories: coordinate regression and heatmap regression. A coordinate regression network performs well on a dataset with sparse landmarks, but not as well as heatmap regression on dense landmarks. However, heatmap regression has been proved that the prediction can be worsen despite MSE improving during the regression of heatmap matching [26]. Luvizon et al. [21] propose the soft-argmax function to convert heatmaps to coordinates to make the network differentiable. Nibali et al. [26] use a new regularisation strategy to improve the prediction accuracy of a network.

2.3 Loss Function

For a CNN-based facial landmark detector, a loss function has to be defined to supervise the network training process. Most existing facial landmark detection approaches are based on the L2 loss, which is sensitive to outliers. Feng et al. [8] propose a new loss function, i.e. wing loss, to balance the sensitivity of small errors and big errors for the training of a deep CNN model. Guo et al. [11] introduce a loss that can adjust weights for different samples during the training process according to the tag that describes the pose of each sample. Merget et al. [23] proposes a loss function that judges whether each landmark is labelled and within the image boundary at the beginning, which gives each landmark a specific weight according to the judgement.

3 Methodology

3.1 Data Augmentation

Data imbalance is a common issue in deep learning, which limits the accuracy and robustness of a trained network [11]. From Table 1 and Fig. 1, we can see that most datasets contain a large number of frontal faces, but lack of samples with large poses, expressions, illuminations and occlusions [42]. The imbalance of a dataset in gesture is very significant. If we train a network using an imbalanced dataset, the network may not able to generalise well to practical applications. Besides, the distribution variations among training and test sets can influence the performance of a trained network significantly.

Table 1. Distribution of the 300-W dataset in gesture [29].

Pose	−30°:−15°	−15°:0°	0°:15°	15°:30°
Pitch	14.59%	61.05%	24.02%	0.34%
Yaw	19.11%	26.02%	22.00%	32.87%

To address the data imbalance problem, various algorithms have been proposed, including both geometric and textural transformations [9]. The main methods used for geometric transformation are flipping, scaling, translation and rotation. For textural transformation, Gaussian noise and brightness transformation are widely used. Nevertheless, if we randomly apply the above methods to the training samples of a dataset, we don't know how many times a training sample should be augmented/copied.

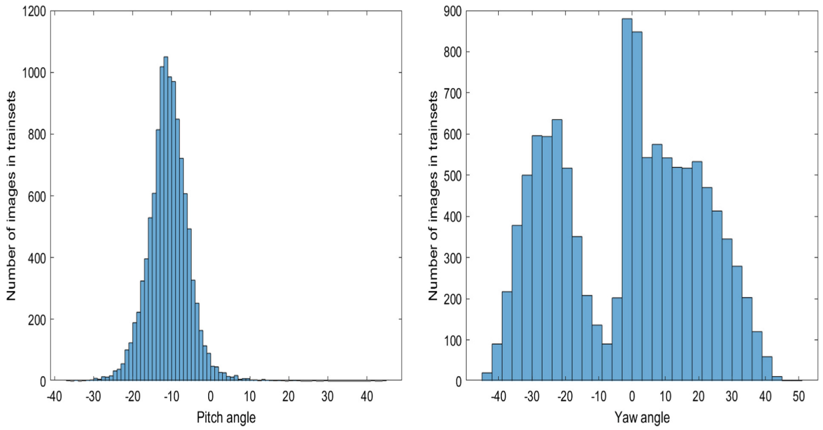

Fig. 1. Distribution of the ICME2019 GC facial landmark datasets in gesture [19]. The X-axis stands for pitch angle in the left figure and yaw angle in the right figure. The Y-axis denotes the number of samples of the training set.

To improve the balance of a dataset, we introduce the Pose-based Data Balancing (PDB) [8] strategy (Algorithm 1) in our work. PDB is a statistical method that aims to analyse the distribution of a face dataset in shape and posture. To adapt PDB to our network, first, we take Procrustes Analysis [10] to align all the faces in a training dataset to the mean face. Procrustes Analysis learns a affine transformation from a shape to another shape with minimum mean square error. By applying PCA to the training set and analysing the distribution of the principle component, we can balance the training set by copying each sample for a fixed number of times which is set to balance the distribution.

Algorithm 1. Pose-based data balancing

```
Require:
    All the images in a training set, $I_n$;
    Bounding boxes of the faces, $B_n$;
    $N * 2$ coordinates of the facial landmarks, $\{(x_n, y_n)\}$;
Ensure:
    Images after copying and random transformation of the training set, $E_n$;
    Bounding boxes of faces after PDB, $B_n$;
    Landmark coordinates after PDB $(x_n, y_n)$
1: Read all the samples in the training set $D_n = \{I_n, B_n, \{(x_n, y_n)\}\}$
2: Calculate the distribution of the training set by Procrustes Analysis;
3: Divide the training set according to the interval of the principal component;
4: Estimate copying times of each image and extend the dataset;
5: Perform random transformation on the samples obtained in step 5;
6: return $D'_n = \{I'_n, B'_n, \{(x'_n, y'_n)\}\}$;
```

In order to minimise the impact of dataset imbalance on facial landmark detection accuracy, the PDB process is applied in each epoch at the beginning. Since the modification of each sample is random, the online PDB process can substantially enhance the variety of samples in different attributes. In each epoch, the data is copied the same number of times, but in different epochs, the data is randomly transformed independently. In this case, dataset is invariably expanded by a period of multiple times and each epoch can be regarded as sampling in a large dataset. According to our experiments, the offline data augmentation has a very good improvement on the performance of a detector. When we convert the offline data augmentation to online data augmentation, the performance of a trained facial landmark detector can be further improved. However, it is worth noting that offline data augmentation does not require many CPU resources. If one does not perform multi-thread data augmentation, each online PDB training process needs to multiply the original running time by several times.

3.2 Network and Mixed Loss

The backbone network of our facial landmark detector is shown in Figs. 2 and 3. The network is based on VGG16 [33] + CPM [7,34], which uses first four convolutions of VGG16 to extract coarse feature maps, followed by three stages of

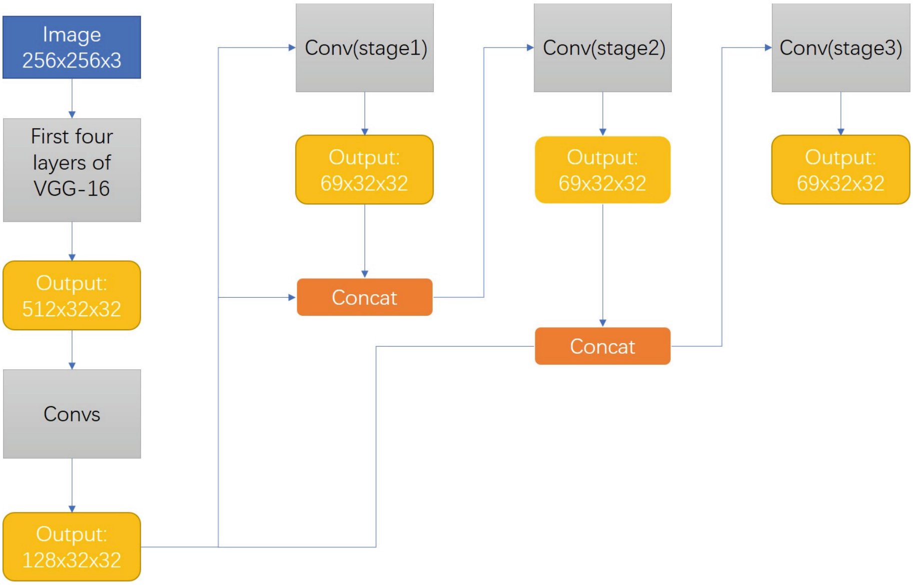

Fig. 2. Backbone network of the proposed facial landmark detector. All the inputs must be resized to 256 × 256. Concat means splicing the feature maps by channels and changing channels by 1 × 1 convolution, the channels 69 in output means 68 landmarks and 1 mask denoting the visibility.

CPM structure. The detailed architecture of conv in Fig. 2 is shown in Fig. 3. We use the convolutional pose machines (CPM) as the main architecture, CPMs combine and concatenate outputs of each stage in the network, in order to hold the geometric constraint and semantic information in feature maps. The ground truth is transformed into heatmap style via taking Gaussian blur on the landmark point. After down-sampling the image with landmark points to the same size and channels with the output of CPM, the error between the predicted and ground truth values is back propagated in each stage of CPM since each stage is intermediately supervised by the L1/L2 loss function.

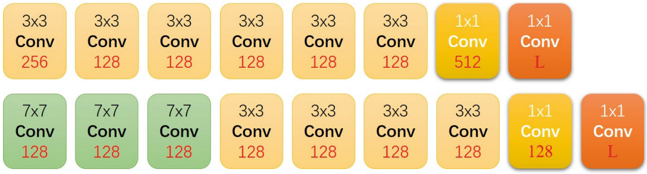

Fig. 3. Detailed kernel size and channels of convolution layers in Conv in Fig. 2, the first row is Conv(stage1), and the second row is the Conv(stage>1). The L in last layers denote the number off facial landmarks and mask.

Models based on heatmap regression have higher accuracy. Additionally, all heatmap regression methods are supervised by the L2 loss, which makes it difficult to improve the form of loss function. However, it is practical to optimise the loss function in coordinates regression, because errors between points are direct. So we try to refine a detector, intending to help the model in learning better parameters by combining coordinate and heatmap regression.

In order to get refined landmark detection in a cascaded model, the multi-stage CPM network is learned in a L2 heatmap regression style. To calculate the loss of the whole network, the multi-stage L2 heatmap loss function and the improved Wing loss function are combined.

$$l_{mix} = \alpha_1 l_{\text{point}} + \alpha_2 l_{stage}, \tag{1}$$

$$l_{stage} = \sum_{i=1}^{3} \beta_i l_{stage}^{(i)}. \tag{2}$$

The form of mixed loss function is shown in (1) and (2), we can see that the network will be updated by point information and heatmap information. The ratio between these two losses are controlled by α_1 and α_2. In heatmap loss, the output of each stage can also contributes to total loss, β_1, β_2 and β_3 are hyper-parameters.

As aforementioned, it is well-known that the proportion of difficult samples is relatively small in a training data set, causing data imbalance issue. Additionally, the simple samples usually dominate the network training. In this case, the widely used L2 loss is not necessarily the best loss function. The L2 loss function amplifies the effects of samples with large errors and neglects small errors. In contrast, the Wing loss function focuses on small and medium errors, but pay less attention to the samples with large errors. In order to design a new loss function that considers the samples with various errors, we formulate the function in Eq. (3) [8]:

$$l_{\text{point}} = wing\left\{x\right\} = \begin{cases} w\ln\left(1+\frac{x}{\varepsilon}\right) & if\ |x|<w \\ |x|-C & otherwise \end{cases}. \tag{3}$$

3.3 Heatmap to Point Regression

It is easy to convert a heatmap to key point coordinates, by just finding the peak locations in the heatmap throughout the argmax function. However, the process is not trivial because the gradients cannot be back-propagated through argmax. To address this issue, this paper adopts soft-argmax, which can guarantee the differentiation in the training process while searching for the maximum value. We represent the argmax function as a parsed form to explain the expectation of the idea. The expectation on the idea of representing the argmax function as a parsed form.

Assuming that one channel of heatmap can be represented as $I(x, y)$, which has the size of $W \times H \times C$, where W and H are the width and height of heatmap, and C denotes channels. The maximum point can be calculated by [26]:

$$softargmax(I) = \left(\sum_{i,j} W_x\left(i,j\right) I\left(i,j\right), \sum_{i,j} W_y\left(i,j\right) I\left(i,j\right)\right), \tag{4}$$

$$W_x\left(i,j\right) = \frac{i}{W}, \tag{5}$$

$$W_y\left(i,j\right) = \frac{j}{H}. \tag{6}$$

In fact, considering that in our model, each heatmap has the order of 10^{-5}, to avoid truncation errors and insufficient precision, we use the adapted Algorithm 2.

Algorithm 2. Modified soft-argmax in our model

```
Input the heatmap;
Set an expansion factor $\alpha$ ;
Take an exponential functional $e^{\alpha x}$ on each item $x$ in the heat map;
Choose a value $x$ in heatmap, and let all the heatmap divides this;
Use the original soft-argmax on the transformed heatmap.
```

4 Experimental Results

4.1 Datasets

In this paper, we conduct experiments on two datasets: the 300-W [29] and AFLW facial landmark datasets [15].

300-W is an open facial landmark dataset, which is composed by LFPW [1], AFW [40], HELEN [18], XM2VTS [24] and IBUG [20] datasets. The whole 300-W dataset contains 3148 training images and 687 test images. Each image in 300-W is labelled with 68 facial landmark (Fig. 4).

AFLW is another classic datasets in face alignment. AFLW consists of more than 25000 images with 21 landmarks. In our experiments, we follow AFLW-Full protocol [15], which contains 24386 images in total, 20000 images for training and others for testing. The images are annotated with 19 landmarks since the landmarks of two ears are ignored in this protocol.

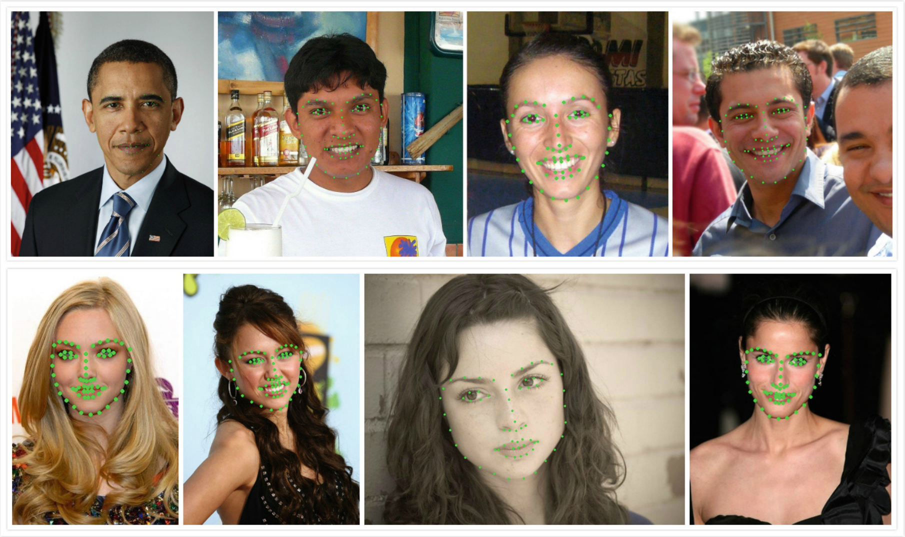

Fig. 4. Partial visualization of the results of our model on 300-W.

4.2 Experimental Settings

We conduct all the experiments on an Intel E5-2650 v4 CPU with two Tesla v100 GPUs. The proposed method was implemented with Pytorch 1.1 [27,28] and Python 3.7. All the input images are resized to $256 \times 256 \times 3$ and the output is N*2 landmark coordinates. The type of heatmap is Gaussian. Our models is updated by Stochastic Gradient Descent (SGD), with the momentum of 0.9 and weight decay of 0.0005. For the 300-W dataset, the learning rate is 0.00005, while for ALFW we set the learning rate to 0.00001. From epoch 30 to 40, the learning

Table 2. Results on 300-W and AFLW datasets. For 300-W, we use inter-pupil distance to compute NME. For AFLW, we use the face size for NME.

Methods	Common set	Challenging set	Full set	AFLW
RCPR [2]	6.18	17.26	8.35	5.43
CFAN [37]	5.50	16.78	7.69	–
TCDCN [38]	4.80	8.60	5.54	–
RAR [35]	4.12	8.35	4.94	–
3DDFA [41]	6.15	10.59	7.01	–
LBF [39]	4.95	11.98	6.32	4.25
ERT [14]	–	–	6.40	4.35
SDM [36]	5.57	15.40	7.50	4.05
Baseline (heatmap + L2 loss)	4.40	9.92	5.49	1.93
Baseline + mixed loss	4.35	9.12	5.43	1.82
Baseline + offline PDB	4.32	8.67	5.38	1.67
Baseline + mixed loss + online PDB	4.26	8.11	5.02	–

rate will decay by a factor of 0.2. After 40 epochs, the learning rate will decay by a factor of 0.1. We train the model for more than 60 epochs. The batch size is set to 64. In the mixed loss function, we try to combine different coefficients with grid search. We get the best result when setting $\alpha_1 = 0.7$ and $\alpha_1 = 0.3$. Meanwhile β_i are set as $\{0.5, 0.5, 1\}$. In the training step, it cost above half a day on 300-W without PDB while about one day with offline PDB. When applying online PDB, it costs 8 days on the same CPU and GPU. For the AFLW dataset we don't do online PDB due to the time limitation.

4.3 Results

We use the backbone network with L2 point regression as the baseline method. Then we try to observe the effects of different methods on 300-W, we use NME as the evaluation metric, which is defined as:

$$NME = \frac{1}{N} \sum_{k=1}^{N} \frac{\|x_k - y_k\|_2}{d}. \tag{7}$$

where x denotes the ground truth landmarks for a given face, y denotes the corresponding prediction and d can be computed as the face size, using the inter-ocular distance or the pupil distance.

4.3.1 Results on 300-W

We apply different innovations to our experiments on 300-W. The performance of different state-of-the-art methods as well as the proposed method in terms of

NME are reported in Table 2. We can see that, in spite of the accuracy loss of point regression our method achieves competitive result. The test batch size is 16 and the proposed method achieves 85 FPS on a Tesla v100 GPU.

4.4 Results on AFLW

As shown in Table 2, we conduct similar experiments on the AFLW dataset. The speed of the proposed method can also achieve more than 80 FPS under the same environment. We also summarise the important parameters, e.g. model size, FLOP and so on. The model size is similar to the model used for 300-W except for the last output layer. The number of parameters is 15.94 M, model size is 127 MB, and the GFLOPs is 2.57 billions.

5 Conclusion

In this paper, we presented a robust facial landmark detector that combines coordinate and heatmap information, thus improving the performance of a trained CNN network in terms of accuracy. Besides, we used the soft-argmax instead of argmax as well as online PDB for training data augmentation. The main purpose of the proposed method is to mitigate the dataset imbalance problem. In addition, we designed a mixed loss function consisting of more information for network training. The experiments obtained on 300-W and AFLW demonstrate the effectiveness of the proposed method compared with the state-of-the-art approaches.

Acknowledgement. This work is partly supported by the National Natural Science Foundation of China (61773117, 61703096 and 61473086), the Jiangsu key R&D plan (BE2017157) and the Natural Science Foundation of Jiangsu Province (BK20170691).

References

1. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 2930–2940 (2011)
2. Burgos-Artizzu, X.P., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion. In: 2013 IEEE International Conference on Computer Vision, pp. 1513–1520 (2013)
3. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1407, pp. 484–498. Springer, Heidelberg (1998). https://doi.org/10.1007/BFb0054760
4. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Comput. Vis. Image Underst. **61**, 38–59 (1995)
5. Cootes, T.F., Walker, K.N., Taylor, C.J.: View-based active appearance models. Image Vision Comput. **20**, 657–664 (2000)
6. Dong, X., Yan, Y., Ouyang, W., Yang, Y.: Style aggregated network for facial landmark detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 379–388 (2018)

7. Dong, X., Yu, S.-I., Weng, X., Wei, S.-E., Yang, Y., Sheikh, Y.: Supervision-by-registration: an unsupervised approach to improve the precision of facial landmark detectors. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 360–368 (2018)
8. Feng, Z.-H., Kittler, J., Awais, M., Huber, P., Wu, X.: Wing loss for robust facial landmark localisation with convolutional neural networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2235–2245 (2018)
9. Feng, Z.-H., Kittler, J., Xiaojun, W.: Mining hard augmented samples for robust facial landmark localization with CNNs. IEEE Signal Process. Lett. **26**(3), 450–454 (2019)
10. Gower, J.C.: Generalized procrustes analysis. Psychometrika **40**, 33–51 (1975)
11. Guo, X., et al.: PFLD: a practical facial landmark detector. ArXiv, abs/1902.10859 (2019)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
13. Kahraman, F., Gökmen, M., Darkner, S., Larsen, R.: An active illumination and appearance (AIA) model for face alignment. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–7 (2007)
14. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1867–1874 (2014)
15. Köstinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: 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), pp. 2144–2151 (2011)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**, 84–90 (2012)
17. Kumar, N., Belhumeur, P., Nayar, S.: FaceTracer: a search engine for large collections of images with faces. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5305, pp. 340–353. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88693-8_25
18. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7574, pp. 679–692. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33712-3_49
19. Liu, Y., et al.: Grand challenge of 106-point facial landmark localization. ArXiv, abs/1905.03469 (2019)
20. Luo, B., Shen, J., Wang, Y., Pantic, M.: The iBUG eye segmentation dataset. In: ICCSW (2018)
21. Luvizon, D.C., Tabia, H., Picard, D.: Human pose regression by combining indirect part detection and contextual information. CoRR, abs/1710.02322 (2017)
22. Lv, J.-J., Shao, X., Xing, J., Cheng, C., Zhou, X.: A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3691–3700 (2017)
23. Merget, D., Rock, M., Rigoll, G.: Robust facial landmark detection via a fully-convolutional local-global context network. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 781–790 (2018)
24. Messer, K., Matas, J., Kittler, J., Luettin, J., Maître, G.: XM2VTSDB: The extended M2VTS database (1999)

25. Milborrow, S., Nicolls, F.: Locating facial features with an extended active shape model. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5305, pp. 504–513. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88693-8_37
26. Nibali, A., He, Z., Morgan, S., Prendergast, L.: Numerical coordinate regression with convolutional neural networks. CoRR, abs/1801.07372 (2018)
27. Paszke, A., et al.: Automatic differentiation in PyTorch, Alban Desmaison (2017)
28. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**, 533–536 (1986)
29. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S.P., Pantic, M.: 300 faces in-the-wild challenge: the first facial landmark localization challenge. In: 2013 IEEE International Conference on Computer Vision Workshops, pp. 397–403 (2013)
30. Saragih, J.M., Goecke, R.: A nonlinear discriminative approach to AAM fitting. In: 2007 IEEE 11th International Conference on Computer Vision, pp. 1–8 (2007)
31. Benitez-Quiroz, C.F., Srinivasan, R., Martínez, A.M.: EmotioNet: an accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5562–5570 (2016)
32. Taigman, Y., Yang, M.W., Ranzato, M., Wolf, L.: DeepFace: closing the gap to human-level performance in face verification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708 (2014)
33. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556 (2015)
34. Wei, S.-E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724–4732 (2016)
35. Xiao, S., Feng, J., Xing, J., Lai, H., Yan, S., Kassim, A.: Robust facial landmark detection via recurrent attentive-refinement networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 57–72. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_4
36. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 532–539 (2013)
37. Zhang, J., Shan, S., Kan, M., Chen, X.: Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8690, pp. 1–16. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10605-2_1
38. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 94–108. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_7
39. Zhu, S., Li, C., Loy, C.C., Tang, X.: Unconstrained face alignment via cascaded compositional learning. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3409–3417 (2016)
40. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2879–2886 (2012)
41. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: a 3d solution. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 146–155 (2016)

42. Zhu, X., Lei, Z., Yan, J., Yi, D., Li, S.Z.: High-fidelity pose and expression normalization for face recognition in the wild. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 787–796 (2015)
43. Liu, F., Zeng, D., Zhao, Q., Liu, X.: Joint face alignment and 3D face reconstruction. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 545–560. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_33
44. Liu, F., Zhao, Q., Liu, X., Zeng, D.: Joint face alignment and 3d face reconstruction with application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(6), 1312–1320 (2017)
45. Lu, J., Liong, V.E., Zhou, X., Zhou, J.: Learning compact binary face descriptor for face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**, 2041–2056 (2015)
46. Lu, J., Tan, Y.-P., Wang, G.: Discriminative multimanifold analysis for face recognition from a single training sample per person. In: 2011 International Conference on Computer Vision, pp. 1943–1950 (2011)