FLASH: Facial Landmark Detection Using Active Shape Model and Heatmap Regression

Nguyen Van Nam[1,2(✉)] and Ngo Thi Ngoc Quyen[3]

[1] Viettel Information Technology, Viettel Group, 7 Alley, TonThatThuyet Street, CauGiay, Hanoi, Vietnam
namnv78@viettel.com.vn
[2] Thuyloi University, 175 Tayson, DongDa, Hanoi, Vietnam
nvnam@tlu.edu.vn
[3] Viettel Cyberspace Center, Viettel Group, 7 Alley, TonThatThuyet Street, CauGiay, Hanoi, Vietnam
quyenntn3@viettel.com.vn

Abstract. Detection of facial landmarks is a critical task for human face identification, emotion recognition in autopilot and real-time visual monitoring applications. This is really challenging due to the high number of discrete landmarks spreading over the face which is of different shapes and may be occluded or obscured. Many methods have been proposed over the years including ASMNet and AnchorFace. However, their performance is still limited in terms of both accuracy and efficiency. In this paper, we propose a novel method for facial landmark detection based on active shape model and heatmap called FLASH. The heatmap aims to highlight the important landmarks. Meanwhile, the shape model helps to conform the distribution of such landmarks. FLASH has been evaluated on two public datasets 300W-Challenging, WFLW and achieved a normalized mean square error (NME) of 6.67%, 7.34% correspondingly, which outperforms most existing methods. Specifically, this is much better than the recent ASMNet method with a NME of 8.20%, 10.77% on the two datasets, respectively. This is also comparable to the state of the art AnchorFace with a NME of 6.19%, 4.62%, correspondingly. The source code of FLASH is also publicly available.

Keywords: facial landmarks · heatmap regression · shape fitting · coordination regression

1 Introduction

In many real-time driver monitoring systems (DMS), facial emotion recognition from images captured by camera is an essential task. In fact, this helps to minimize potential accidents caused by attentionless drivers by detecting if there exists any sleepy, drunk or tired expressions in their face and notifying

N.-S. Vo and H.-A. Tran (Eds.): INISCOM 2023, LNICST 531, pp. 170–182, 2023.
https://doi.org/10.1007/978-3-031-47359-3_13

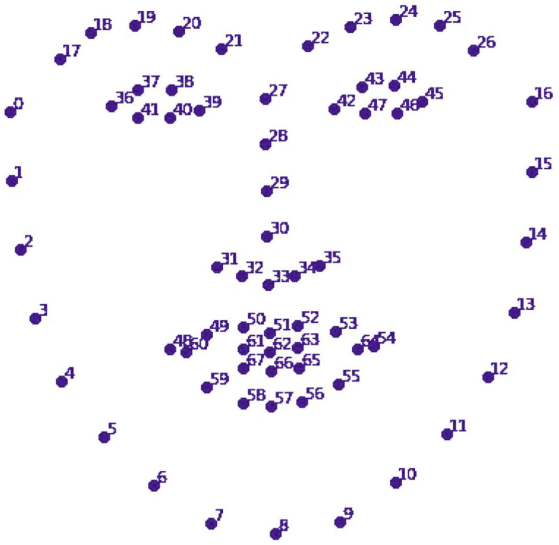

Fig. 1. Description of 68 points of facial landmark.

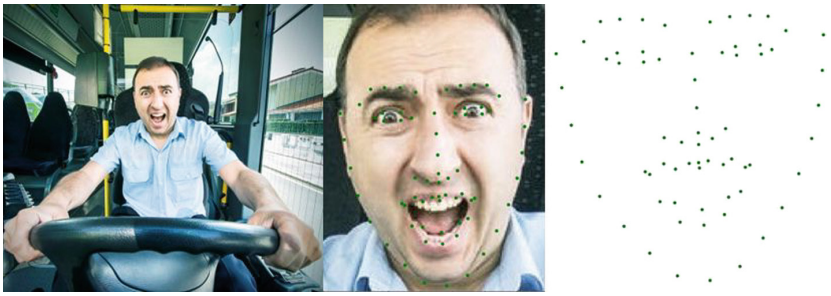

Fig. 2. Demonstration of the detection of 68 points of facial landmark on a given image: The left one is of a bus driver. The middle one denotes his angry face associated with 68 points detected. The right one describes the resulted locations of the landmarks on an angry face.

them accordingly. Existing methods for facial emotion recognition usually take as input either the hidden features or the transparent landmarks of the facial images. However, the latter ones are preferable thanks to its explainability.

Facial landmarks include certain principal points locating around the eyes, the nose, the mouth and the boundary of a human face as shown in Fig. 1. The task of facial landmark detection is to locate these points in a given face image as depicted in Fig. 2[1]. This problem is very challenging due to the high number

[1] The original image is referred from https://www.dreamstime.com/photos-images/bus-driver.html.

of discrete points distributed in various shapes of face which can be captured from different angles of view.

Earlier efficient approaches for facial landmark detection are of shape fitting models including Active Shape Model (ASM) [6], Active Appearance Models (AAM) [5], Constrained Local Model (CLM) [7], Discriminative Response Map Fitting (DRMF) [1] as well as DeFA [20]. However, their efficiency is only limited on individual datasets due to the fact that their final prediction depends mainly on the initial shape. In fact, such methods are fast converged but not well generalized.

Recent deep learning models usually consist of a convolutional neural networks (CNN) backbone and a regression head. The direct regression methods like take as input the last flattened feature map of the CNN backbone. Meanwhile, in heatmap regression models such as Style Aggregation Network (SAN) [9] and MobileFAN [33] certain de-convolutional layers are added to the CNN backbone to form a fully CNN (called FCN). Its outputted set of 2D heatmaps with the same width and height to the original image are then used for locating position of the landmarks. These models are well generalized thanks to their non-linearity but hardly converged due to high number of dispersal landmarks.

In this paper, we propose a novel model for **F**acial **L**andmark detection combining the **A**ctive **S**hape model and the **H**eatmap regression called FLASH. Our main contributions are therefore three-fold. Firstly, we designed a FCN backbone based on Resnet to produce a set of high quality heatmaps. Secondly, we added a softmax-argmax layer to our backbone for locating the position of the landmarks on the corresponding heatmap. Then, we combined the active shape model loss and the heatmap regression loss in an efficient manner. Thirdly, we trained and evaluated FLASH on two public datasets 300W-Challenging and WFLW achieved a normalized mean square error (NMSE) of 6.67% and 7.34%, respectively. These results are much better than DeFA (9.38%), MobileFAN (6.87%) and SAN (6.60%) on the 300W-Challenging dataset. These also overcome the coarse-to-fine face shape searching CFSS [4](9.07%) and the ASM-Net [11] (10.77%) which is a combination of active shape models and a direct regression network on WFLW dataset.

The rest of the paper is organized as follows. In Sect. 2, we study recent approaches for facial landmark detection relating to our work. Then, our proposed model FLASH is presented in detail in Sect. 3. We summarize and analyze the experimental results of FLASH on two public datasets in Sect. 4. Open issues about the work are discussed in Sect. 4.5. Finally, Sect. 5 concludes our works.

2 Related Works

Over decades, lots of methods for facial landmark detection have been proposed. In this paper, we concerns mainly on the visual landmark regression and the facial shape fitting.

2.1 Shape Fitting

Traditional template matching approaches such as ASM [6], AAM [5], CLM [7] and DeFA [20] detect the facial landmarks by learning their common distribution and from a mean shape, computed from certain active samples, regressing them. ASM is based on the dimension reduction method Principle Component Analysis (PCA) [16] for shape fitting. AAM improved the performance of ASM by combining both the shape and appearance models in iterative manner. CLM introduced another appearance sampling technique in which the pixel values in the texture patches are normalized with zero mean and unit variance. Using CNN, DeFA models the facial shape in 3D to not only aligns facial landmarks but also matches SIFT (Scale-Invariant Feature Transform) points as well as the facial contours. However, due to limited feature engineering, the performance of such approaches are limited especially in case of occluded face images.

2.2 Landmark Regression

As introduced, the neural networks for facial landmark detection usually include a CNN backbone and a regression head which is fed with a feature vector. The networks can be categorized as coordinate and heatmap regression according to the way such vector is built from the backbone.

Coordinate Regression. In case of coordinate regression networks, any CNN encoder can be used as their backbone. The regression head is directly fed with the flattened feature embedding of the backbone. Mnemonic Descent Method (MDM) [26] is a combined convolutional recurrent neural network which aims to cooperate the regressors of facial landmarks. DeepReg [21] is a deep regressor for gradual detection of facial landmarks with two-stage initialisation. In Wing [12], the wing regression loss was proposed for landmark localization rather than the L1 and L2 losses thanks to its ability to help the regression networks not only deal with large localization errors as L1 and L2, but treat also well the medium and small localization ones. Wing has been experimented with Resnet-50 [13] backbone. However, such average loss for regression of a high number of positions on the whole face is unable to assure small prediction errors for individual landmarks.

Heatmap Regression. The heatmap regression networks such as AWing [27], MobileFAN [33], Gaussian Vector (GV) [31] and AdNet [15] are autoencoder backbone which is composed of a CNN encoder and a decoder to produce probability distributions in form of heatmaps corresponding to the facial landmarks. In each heatmap, the position with the highest probability is chosen for the respective landmark.

AWing proposed an adaptive Wing loss function for coordinate regression from facial boundary map for better conforming the heatmap pixels to the facial shape. Gaussian Vector (GV) converts heatmap in to a pair of vector for

each landmark to preserve spacial information and simplify the post-processing. AdNet introduced anisotropic direction loss and anisotropic attention module for better learning the facial structure as well as the texture details and mitigating the error-bias of facial landmarks.

2.3 Joint Shape Fitting and Regression Networks

There are also few methods which combine the shape fitting approach and the regression network such as LAB [28], ASMNet [11] and AnchorFace [32]. LAB is a combination of the boundary fitting and the coordinate regression. Using a stacked Hourglass network [24] as an autoencoder backbone to produce facial boundary map, LAB then regresses the coordination of facial landmarks from the boundary in order to avoid the ambiguities of such key-points. ASMNet leveraged the light-weight MobileNetV2 [25] as backbone and presented a multi-task loss which is the sum of the mean square error and the active shape model loss. This enables ASMNet to learn both the shape and the coordination of the facial landmarks with less parameters than LAB.

In AnchorFace, the authors introduced certain anchor templates and regress the offsets on each template. They then aggregates the predictions on every templates to produce the final results. AnchorFace utilized ShuffleNetV2 [22] as its backbone. AnchorFace can deal with face poses of large variations thanks to its anchor templates. Nevertheless, the anchor templates need to be carefully selected and the inference time must be improved. AnchorFace is also known as anchor-based method.

Such joint approaches are usually more performant than the separate ones. However, existing joint methods are only between coordinate regression and the shape fitting. In this paper, we propose FLASH, a facial landmark detection method based on shape fitting and heatmap regression to fill the gap as well as to leverage the robustness of such combination.

3 FLASH: The Proposed Method

Our proposed method FLASH consists of a heatmap regression network a training loss function including both the coordination and the shape matching errors. Two principal components of the heatmap regression network are the heatmap-generated backbone and the heatmap regression head.

3.1 The Heatmap-Generated Backbone

As depicted in Fig. 3, the backbone is an autoencoder which takes as input the face image of size 224×224 and produces a set of heatmaps. The encoder is composed of five multi-filter convolutional layers which are activated by ReLU function and dimensionally reduced by Max-Pooling. Meanwhile, the decoder is based on three deconvolutional layers to produce a set of heatmaps of the same size with the input image, each of which corresponds to a facial landmark.

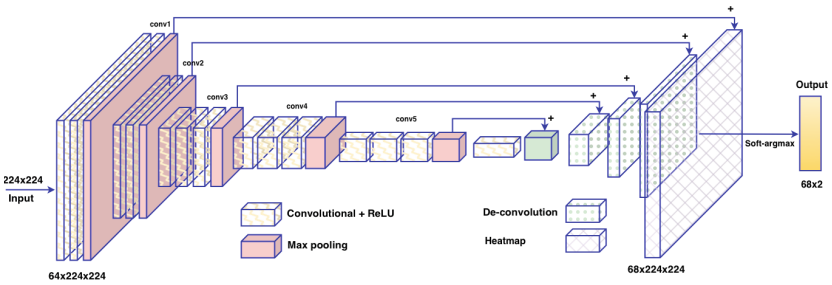

Fig. 3. The network architecture of FLASH.

3.2 The Heatmap Regression Head

Given a set of n heatmaps $H = \{H^i\}, i = \overline{1,n}$, each of size K × K (in this case K is qual to 224) and flattened to a vector of K^2 dimensions $h^i = (h^i_1, h^i_2, .., h^i_{K^2})$, the regression head of FLASH can predict the coordination for the respective facial landmarks using a soft arg-max function as follows:

$$\{\hat{x}_i, \hat{y}_i\} = softargmax_j(j \cdot f(j)) \tag{1}$$

where $f(j), j = \overline{1, K^2}$ is a probability distribution function defined as follows

$$f(j) = \frac{e^{\alpha \cdot h^i_j}}{\sum_{k=1}^{K^2} e^{\alpha \cdot h^i_k}} \tag{2}$$

in which $\alpha \geq 1$ is the temperature parameter. For the i^{th} heatmap H^i, the function $softargmax$ returns an index j^* where $f(j^*)$ is the maximal value of $\{f(j), \forall j = \overline{1, K^2}\}$. From j^*, we can calculate the coordination $(\hat{x}_i, \hat{y}_i)$ for the corresponding i^{th} facial landmark. This function can be differentiated that can be used in FLASH instead of the traditional $argmax$ and $softmax$ functions.

3.3 The Multitask Loss Function

As we aim to integrate the facial landmarks in to a given shape, we designed a multitask loss function for training our proposed network.

The Coordination Loss. The mean square error is used as the coordination loss as follows:

$$\mathcal{L}_{coord} = \frac{1}{n} \sum_{i=1}^{n} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_j)^2] \tag{3}$$

where n is the number of facial landmarks, $(x_i, y_i), (\hat{x}_i, \hat{x}_j), i = \overline{1,n}$ is the ground truth and predicted coordination of the i^{th} facial landmark, respectively.

The Shape Loss. Given a training set with m samples in which the $i^{th}, i = \overline{1,m}$ is represented as a vector of $2n$ dimensions $s^i = (x_1^i, y_1^i, x_2^i, y_2^i, ..x_n^i, y_n^i)$, using PCA (Principal Component Analysis) [6], this can be approximated by $\tilde{s}^i$ as follows:

$$\tilde{s}^i = \overline{s} + P \cdot b^i \tag{4}$$

where $\overline{s}$ is the mean shape

$$\overline{s} = \frac{1}{m} \sum_{j=1}^{m} s^j \tag{5}$$

and $P = (p_1|p_2|..|p_t)$ is a matrix constituted from t eigenvectors with the highest corresponding eigenvalues $\lambda_1, \lambda_1, .., \lambda_t$ of the following co-variance matrix:

$$S = \frac{1}{m-1} \sum_{j=1}^{m} (s^j - \overline{s})(s^j - \overline{s})^T \tag{6}$$

and b^i is a t-dimensional vector containing a set of parameters for a deformable model:

$$b^i = P^T(s^i - \overline{s}) \tag{7}$$

The shape loss is then calculated as follows

$$\mathcal{L}_{shape} = \frac{1}{2 \cdot n} \sum_{j=1}^{2n} (s_j^i - \tilde{s}_j^i)^2 \tag{8}$$

The Multitask Loss. For every training samples, the overall loss is the combination of the coordinate and the shape ones as the following

$$\mathcal{L} = \mathcal{L}_{coord} + \beta \cdot \mathcal{L}_{shape} \tag{9}$$

where β is the shape fitting rate which varies in reverse proportionally to the number of the training epochs for FLASH. This is because as many other convolutional neural networks, FLASH learns the shape before featuring the pixel-wise image. The ratio can then be defined as the following discrete function:

$$\beta = \begin{cases} 2 & \text{if } e \leq \frac{N_e}{5} \\ 1 & \text{if } e \leq 2 \cdot \frac{N_e}{5} \\ 0.5 & \text{if } e \leq 3 \cdot \frac{N_e}{5} \\ 0 & \text{if } e > 3 \cdot \frac{N_e}{5} \end{cases} \tag{10}$$

where e, N_e is the current and total number of training epochs, respectively. At the initial steps of FLASH training where the shape features are important, the shape fitting rate β is also high enough. Reversely, at the final steps, β is set to zero since there exists mainly pixel featuring in the network.

4 Experiments

4.1 Datasets

Our proposed FLASH method is evaluated on two famous facial landmark datasets including 300W and WFLW.

300W. The 300W dataset totally consists of 3837 facial images with 68 landmarks annotated. The training set includes 3148 images in which 2000 are from HELEN [19], 811 from LFPW [2] and 337 from AFW [17]. The full testing set is composed of 689 images which is divided in to a common set of 554 combining those from HELEN and LFPW and a challenging set with 135 images.

WFLW. The WFLW dataset [28] includes 10000 facial images which are annotated by 98 landmarks. Three fourths of the dataset are used for training and the rest for testing. This latter is composed of six subsets with different difficulties including 314 for expression, 326 for large pose, 206 for make-up, 736 for occlusion, 698 for illumination and 773 for blurring.

4.2 Evaluation Metrics

As commonly used for benchmarking of facial landmark detection methods, we also based on the **normalized mean error (NME)** to evaluate the accuracy of our proposed FLASH as follows:

$$NME = \frac{1}{n}\sum_{i=1}^{n}\frac{\sqrt{(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2}}{d} \tag{11}$$

where d is the distance between the two outer eye corners (inter-ocular) specifically for each dataset. This is also the normalized factor used in the 300W and WFLW datasets.

The **failure rate (FR)** is also involved in this case to evaluate the robustness of the methods in term of NME. This indicates the rate of failed recognition in which NME is less than 10%. The smaller FR is, the more powerful the model is.

4.3 Model Training

The input images are all resized to 224×224 before training. FLASH used Resnet 50 as its backbone for better heatmap featuring and is implemented in Pytorch. The model is trained by 50 epochs using Adam optimizer with the learning rate of 10e-5, the decay of 10e-5 and batch size of 64 on a K80 GPU of Google Colaboratory.

4.4 Results

Evaluation Results on 300W Dataset. The results of FLASH on 300W dataset can be seen on the Table 1. Our model FLASH achieved a NME of 3.79%, 6.67% and 4.35% on the Common, Challenging and Full subset of 300W, respectively. These outperform most of the recent methods of coordinate regression, heatmap regression as well as shape fitting such as DeFA, MobileFAN, PCD-CNN, CPM, ASMNet especially on the Challenging subset. FLASH is a bit less accurate than the state-of-the-art AnchorFace but it runs faster at the rate of 43 frames per second (FPS) on NVIDIA Tesla K80 GPU than AnchorFace with 45 FPS on much more powerful NVIDIA GTX Titan Xp GPU. These results prove the efficiency of the combination between the heatmap regression and the shape fitting in our FLASH method.

Table 1. Accuracy of FLASH and other comparative methods on 300W dataset.

Model	Category	Common	Challenging	Full
CFSS [34]	Shape Fitting	4.73	9.98	5.76
DSRN [23]	Coordinate Regression	4.12	9.68	5.21
DeFA	Shape Fitting	5.37	9.38	6.10
RDR [29]	Coordinate Regression and Shape Fitting	5.37	9.38	6.10
RCN [14]	Coordinate Regression	4.67	8.44	5.41
ASMNet	Coordinate regression and Shape Fitting	4.82	8.20	5.50
CPM [10]	Coordinate Regression	3.39	8.14	4.36
PCD-CNN [18]	Heatmap Regression	3.67	7.62	4.44
CPM+SBR [10]	Coordinate Regression	3.28	7.78	4.10
MobileFAN	Heatmap Regression	4.22	6.87	4.74
ODN [8]	Coordinate Regression	3.56	6.67	4.17
SAN	Coordinate Regression	3.34	6.60	3.98
AnchorFace	Anchor-based Regression	**3.12**	**6.19**	**3.72**
FLASH (ours)	Heatmap Regression and Shape Fitting	*3.79*	*6.67*	*4.35*

Evaluation Results on WFLW Dataset. FLASH is also evaluated on the WFLW dataset using both NME and FR metrics as in Table 2. FLASH achieved the best performance and robustness with a NME of 7.34% and a FR of 17.08% in comparison with recent advanced methods such as ESR (with NME of 11.13%, FR of 35.24%), SDM (with NME of 10.29%, FR of 29.40%), CFSS (with NME of 9.07%, FR of 20.56%) and ASMNet (with NME of 10.77%, FR of 39.12%) on the full WFLW dataset. However, these results are far from those of AnchoFace with NME of 4.62% and FR of 4.2% on the full dataset. This is because FLASH is not efficient for the large pose, occlusion and blur subsets with a NME of 14.81%, 9.10%, 8.15% and a FR of 64.11%, 25.95% and 19.40%, respectively. In

fact, AnchorFace is fine-tuned according to various shapes while our FLASH is relied on only one for a given dataset.

Table 2. Accuracy of FLASH and other comparative methods on WFLW dataset.

Data	Metric	ESR [3]	SDM [30]	CFSS	ASMNet	AnchorFace	FLASH (ours)
Full	NME	11.13	10.29	9.07	10.77	**4.62**	*7.34*
	FR	35.24	29.40	20.56	39.12	**4.2**	*17.08*
Large Pose	NME	25.88	24.10	21.36	21.11	–	**14.81**
	FR	90.18	84.36	66.22	98.41	–	**64.11**
Expression	NME	11.47	11.45	10.09	12.02	–	**7.74**
	FR	42.04	33.44	23.25	59.87	–	**14.33**
Illumination	NME	10.49	9.32	8.30	9.93	–	**6.92**
	FR	30.80	26.22	17.34	33.38	–	**12.75**
Makeup	NME	11.05	9.38	8.74	10.55	–	**7.16**
	FR	38.84	27.67	21.84	38.34	–	**16.50**
Occlusion	NME	13.75	13.03	11.76	12.34	–	**9.10**
	FR	47.28	41.85	32.88	48.64	–	**25.95**
Blur	NME	12.20	11.28	9.96	11.62	–	**8.15**
	FR	41.40	35.32	23.67	46.31	–	**19.40**

4.5 Discussion

Facial landmark detection is an active research topic over many years because this can be more efficiently used to recognize the human facial emotion than relying on the whole human face. However, most recent methods focus more on the feature engineering of the individual facial landmarks but less on their distribution meaning the shape of the face. Although, the power of deep learning backbone networks has been thoroughly leveraged, the performance of such coordination and heatmap regression methods remains limited. ASMNet was the first to take in to account the shape fitting in to its coordination regression and initially gained positive results. However, the coordination regression approach aims to extract features at the cell level while the heatmap regression targets to the pixel level of the image which is closer to the facial landmarks in this case. Our proposed method FLASH is a combination of heatmap regression and shape fitting achieved a much better performance and robustness than ASMNet in both 300W and WFLW datasets which proved our judgments.

5 Conclusion

As discussed, the facial landmark detection is necessary for recognition of human emotion which can be applied in advanced driver assistance systems. This task

is really hard due to the dispersion of high number of landmarks on the human face. Efficient methods such as ASMNet and AnchorFace all take in to account their distribution meaning the facial shape. However, these coordination regression methods extract the feature at the cell level which is less accurate than at the pixel level as in case of heatmap regression. In this paper, we proposed a novel facial landmark detection method called FLASH which is the first combination between heatmap regression and shape fitting. The evaluation on 300W and WFLW datasets showed that FLASH outperforms many existing methods including ASMNet. FLASH can not be compared to AnchorFace due to using less number of anchor shapes. These results proved that such combination is reasonable and the FLASH can also be better improved with more performant backbone and more facial priors.

References

1. Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Robust discriminative response map fitting with constrained local models. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3444–3451. CVPR 2013, IEEE Computer Society, USA (2013). https://doi.org/10.1109/CVPR.2013.442
2. 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**(12), 2930–2940 (2013). https://doi.org/10.1109/TPAMI.2013.23
3. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2887–2894 (2012). https://doi.org/10.1109/CVPR.2012.6248015
4. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. Int. J. Comput. Vision **107**(2), 177–190 (2014)
5. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: Burkhardt, H., Neumann, B. (eds.) Computer Vision – ECCV 1998, pp. 484–498. Springer, Berlin Heidelberg, Berlin, Heidelberg (1998)
6. Cootes, T., Baldock, E., Graham, J.: An introduction to active shape models. Image Process. Anal. **328**, 223–248 (2000)
7. Cristinacce, D., Cootes, T.: Feature detection and tracking with constrained local models, vol. 41, pp. 929–938 (2006). https://doi.org/10.5244/C.20.95
8. Ding, H., Zhou, P., Chellappa, R.: Occlusion-adaptive deep network for robust facial expression recognition. In: 2020 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–9. IEEE Press (2020). https://doi.org/10.1109/IJCB48548.2020.9304923
9. Dong, X., Yan, Y., Ouyang, W., Yang, Y.: Style aggregated network for facial landmark detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 379–388 (2018)
10. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 360–368 (2018)

11. Fard, A.P., Abdollahi, H., Mahoor, M.H.: ASMNet: a lightweight deep neural network for face alignment and pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19–25, 2021, pp. 1521–1530. Computer Vision Foundation / IEEE (2021). https://doi.org/10.1109/CVPRW53098.2021.00168
12. Feng, Z., 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 (CVPR), pp. 2235–2245. IEEE Computer Society, Los Alamitos, CA, USA (2018). https://doi.org/10.1109/CVPR.2018.00238, https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00238
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. CVPR 2016, IEEE (2016). https://doi.org/10.1109/CVPR.2016.90, http://ieeexplore.ieee.org/document/7780459
14. Honari, S., Yosinski, J., Vincent, P., Pal, C.: Recombinator networks: learning coarse-to-fine feature aggregation. In: Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on. IEEE (2016)
15. Huang, Y., Yang, H., Li, C., Kim, J., Wei, F.: ADNet: leveraging error-bias towards normal direction in face alignment. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3060–3070 (2021)
16. Jolliffe, I.T., Cadima, J.: Principal component analysis: a review and recent developments. Philos. Trans. Royal Soc. A Math. Phys. Eng. Sci. **374**(2065), 20150202 (2016)
17. 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). https://doi.org/10.1109/ICCVW.2011.6130513
18. Kumar, A., Chellappa, R.: Disentangling 3D pose in a dendritic CNN for unconstrained 2D face alignment. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 430–439. IEEE Computer Society, Los Alamitos, CA, USA (2018). https://doi.org/10.1109/CVPR.2018.00052, https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00052
19. 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.) Computer Vision - ECCV 2012, pp. 679–692. Springer, Berlin Heidelberg, Berlin, Heidelberg (2012)
20. Liu, Y., Jourabloo, A., Ren, W., Liu, X.: Dense face alignment. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 1619–1628 (2017)
21. Lv, 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). https://doi.org/10.1109/CVPR.2017.393
22. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: ShuffleNet V2: practical guidelines for efficient CNN architecture design. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018, pp. 122–138. Springer International Publishing, Cham (2018)
23. Miao, X., Zhen, X., Liu, X., Deng, C., Athitsos, V., Huang, H.: Direct shape regression networks for end-to-end face alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)

24. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision - ECCV 2016, pp. 483–499. Springer International Publishing, Cham (2016)
25. Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.: MobileNetV2: inverted residuals and linear bottlenecks. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018, pp. 4510–4520. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00474
26. Trigeorgis, G., Snape, P., Nicolaou, M.A., Antonakos, E., Zafeiriou, S.: Mnemonic descent method: a recurrent process applied for end-to-end face alignment. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4177–4187. IEEE Computer Society, Los Alamitos, CA, USA (2016). https://doi.org/10.1109/CVPR.2016.453, https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.453
27. Wang, X., Bo, L., Fuxin, L.: Adaptive wing loss for robust face alignment via heatmap regression. In: The IEEE International Conference on Computer Vision (ICCV) (2019)
28. Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at boundary: a boundary-aware face alignment algorithm. In: CVPR (2018)
29. Xiao, S., et al.: Recurrent 3D–2D dual learning for large-pose facial landmark detection. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 1642–1651 (2017). https://doi.org/10.1109/ICCV.2017.181
30. 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). https://doi.org/10.1109/CVPR.2013.75
31. Xiong, Y., Zhou, Z., Dou, Y., Su, Z.: Gaussian vector: an efficient solution for facial landmark detection. In: Ishikawa, H., Liu, C.-L., Pajdla, T., Shi, J. (eds.) ACCV 2020. LNCS, vol. 12626, pp. 70–87. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-69541-5_5
32. Xu, Z., Li, B., Yuan, Y., Geng, M.: AnchorFace: an anchor-based facial landmark detector across large poses. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 3092–3100 (2021)
33. Zhao, Y., Liu, Y., Shen, C., Gao, Y., Xiong, S.: MobileFAN: transferring deep hidden representation for face alignment. Pattern Recognit. **100**, 107114 (2019). https://doi.org/10.1016/j.patcog.2019.107114
34. Zhu, S., Li, C., Loy, C.C., Tang, X.: Face alignment by coarse-to-fine shape searching. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4998–5006 (2015). https://doi.org/10.1109/CVPR.2015.7299134