An improved boundary-aware face alignment using stacked dense U-Nets

Jianghao Ye, Ying Cui, Xiang Pan, Herong Zheng, Dongyan Guo and Yanan Ren

Abstract
Facial landmark localization is still a challenge task in the unconstrained environment with influences of significant variation conditions such as facial pose, shape, expression, illumination, and occlusions. In this work, we present an improved boundary-aware face alignment method by using stacked dense U-Nets. The proposed method consists of two stages: a boundary heatmap estimation stage to learn the facial boundary lines and a facial landmark localization stage to predict the final face alignment result. With the constraint of boundary lines, facial landmarks are unified as a whole facial shape. Hence, the unseen landmarks in a shape with occlusions can be better estimated by message passing with other landmarks. By introducing the stacked dense U-Nets for feature extraction, the capacity of the model is improved. Experiments and comparisons on public datasets show that the proposed method obtains better performance than the baselines, especially for facial images with large pose variation, shape variation, and occlusions.

Keywords
Face alignment, heatmap, boundary-aware, dense U-Nets

Introduction
Facial shape that composed of a series of corresponding facial landmarks is an important face description information for face analysis. Facial landmark localization, as so called face alignment, plays a key role in various applications, such as face recognition, facial expression recognition, facial attribute analysis, and so on. While extensive efforts have been devoted and large step has been made in recent years, the performance of existing face alignment methods still unsatisfactory in real tasks. Faces captured in real world with unconstrained conditions usually suffer from significant variations of pose, expression, illumination, and partial occlusions.

In recent years, deep neural networks have been widely applied in the fields of image understanding and computer vision. By formulating face alignment as a heatmap regression problem, these methods aim to learn a set of landmark heatmaps, which illustrate the probability of the landmark location at one pixel. Since heatmap regression is a dense regression problem, the model can be improved by studying the feature representation between resolutions and connecting the spatial information of each resolution. Among various heatmap regression methods,
the stacked hourglass network\textsuperscript{8} which can extract the deep features from the original scale without changing the dataset size has recently achieved great success. Inspired by the stacked hourglass networks, Guo et al.\textsuperscript{9} propose a stacked dense U-Nets which improve the model capacity without increasing the computational complexity by increasing the model size and simplifying the sampling step. The dual transformers of inside and outside make the stacked dense U-Nets spatially invariant to an arbitrary face, which make the method much more robust for face alignment in the wild condition. However, the lack of explicit constraint among landmarks makes the method still cannot well deal with facial images with large shape variation and occlusions.

Since most of the facial landmarks are not corner points, their semantical location may change greatly with large shape variation, pose variation, and occlusion, which make them difficult to be localized. Many methods represent the geometric structure of the face by constructing the relationship among landmarks implicitly or explicitly. The classical active appearance model (AAM)-based methods\textsuperscript{11,12} and cascaded regression model-based methods\textsuperscript{13–15} generally treat the face landmarks as a whole shape to remain the structure. For deep neural networks, Wu et al.\textsuperscript{16} introduce a boundary-aware regressor to represent the structure by using well-defined facial boundaries. The boundary heatmaps algorithm can make good use of facial contour information.

In this work, we take advantage of the stacked dense U-Nets for improving the capacity of feature extraction model and the boundary heatmaps for facial shape structure constraint, proposing an improved boundary-aware face alignment method by using stacked dense U-Nets. The boundary line of human face can be used as the key point location medium, and the stacked dense U-Nets can extract the feature information of different scales to improve the effect of heatmaps regression. The proposed method is evaluated on three popular benchmarks including WFLW,\textsuperscript{16} 300-W,\textsuperscript{17} and COFW.\textsuperscript{18} Experimental results demonstrate that our method has a good performance on challenging faces with occlusions, large pose variation, and shape variation. The accuracy of landmarks on the face contour has been greatly improved.

**Related works**

As a fundamental topic in face recognition area, face alignment has received considerable attention for a long period. The popular face alignment methods have revolved around from active shape model,\textsuperscript{19} AAM,\textsuperscript{11} moving to Gauss–Newton deformable part model,\textsuperscript{20} constrained local model,\textsuperscript{12} and cascaded regression models.\textsuperscript{13–15} The current state-of-the-art methods revolve around deep convolutional neural networks, especially the heatmap regression models.\textsuperscript{21–24} In the following, we briefly introduce the recent face alignment methods with a special focus on heatmap regression models.

The heatmap regression models can generate a heatmap for each key point, which can be used to estimate the current landmark location in the image. In the deep alignment network (DAN),\textsuperscript{24} DAN is divided into multiple stages, and the landmark heatmaps are used as the input of the intermediate stage to transfer the information about estimated landmark location in the previous stage. Yang et al.\textsuperscript{22} use the supervised transformations to normalize faces and obtain prediction heatmaps through a stacked hourglass network.\textsuperscript{8} Deng et al.\textsuperscript{25} propose a joint multiview convolutional network for faces in-the-wild with large pose variations and achieve multiview face alignment by using a stacked hourglass network.\textsuperscript{8} Merget et al.\textsuperscript{21} generate heatmaps based on the labeled dataset and Gaussian model and then use the heatmap as a guide for supervised learning with convolutional neural network to estimate the facial landmarks. Inspired by the efficiency of boundary detection in vision tasks, Wu et al.\textsuperscript{16} propose to estimate facial boundary heatmaps by using the stacked hourglass structure\textsuperscript{8} and then utilize the boundary lines which identifies the facial geometry structure to help face alignment.

Hourglass network with symmetrical structure\textsuperscript{8} is proposed to capture information at each scale in human pose estimation, and it also performs well in face alignment\textsuperscript{25}. Chu et al.\textsuperscript{26} use the stacked hourglass network to obtain different levels of heatmaps by changing the inputs between structures. Yang et al.\textsuperscript{27} use the stacked hourglass network and generate a score map of joint positions at the end of each hourglass. To improve model capacity, Guo et al.\textsuperscript{9} propose a scale aggregation topology (SAT) by adding down-sampling inputs for aggregation nodes. They present a channel aggregation block which increases robustness when local observation is blurred. Compared with the original hourglass network, the dense U-Net greatly improves the capacity of the model while maintaining similar computational complexity and model size.

**Proposed method**

In this section, we describe the proposed face alignment framework in details. As shown in Figure 1(a) and (b), the proposed framework consists of two stages: (a) boundary heatmap estimation stage to generate boundary lines for the later stage assistance; (b) boundary-aware facial landmark localization stage to predict the final face alignment result with the incorporated facial boundary information.

The facial boundaries as the representation of facial geometric structure can present accurate and universal geometric structure of the face. The facial boundary heat map can be derived from the facial landmarks of different annotations as constraints for the next regression. Since the quality of boundary heatmaps is crucial for the final landmark regression, following the work boundary heatmap estimation,\textsuperscript{16} an adversarial learning\textsuperscript{28} operation between
the estimated boundary heatmaps and the ground-truth boundary heatmaps is added to discriminate effective boundaries, as shown in Figure 1(c). In the facial landmark localization stage, we use the estimated boundary heatmaps as a constraint condition into the stacked dense U-Nets to predict the landmarks.

Boundary heatmap estimation
Facial boundary is closely related to the facial landmark. More than piecemeal landmarks, facial boundary can well describe the geometry structure of a face. When associate each facial landmark with a semantical boundary, such as the mouth and the nose bridge, we can get a whole shape structure that covers the face. In this manner, the piecemeal landmarks are related to each other by the geometrical and semantical relationship among them. Therefore, when large occlusion or large pose variation occurred, the influenced landmarks can be better estimated by the current related stable landmarks through a message passing strategy. Motivated by that, we aim to estimate boundary heatmaps to obtain facial boundary information.

In order to fuse boundary lines in feature learning, we need to define facial boundary heatmaps to aid the learning of feature. Given a facial image $I$ and $n$ corresponding landmark annotations $S_l$, $l = 1 : n$, we can define $K$ subsets $S_i \subset S$ to represent the boundary markers belonging to $K$ boundaries, such as chin and lower lip. A complete boundary line can be obtained by interpolating $S_i$ in each boundary. Setting the point on the complete boundary line to 1, and the others to 0. Then the binary boundary map $B_i$ can be formed of the same size as $I$. And the distance map $D_i$ can be obtained by distance transformation of the binary boundary map $B_i$. Finally, a Gaussian expression is used to transform the distance map $D_i$ to real boundary heat map $M_i$ which we need.

To implement the boundary heatmap estimation, here we utilize the boundary heatmap estimator proposed by LAB $^{16}$ as the baseline, and the mean squared error (MSE) between the ground-truth and estimated boundaries is used as the loss function. Different from LAB, here we use the superimposed SAT $^9$ as the backbone instead of hourglass network. Compared with hourglass network, SAT has better advantages in feature extraction and model capacity.

As shown in Figure 1(a), the input face image is put into the stacked U-Nets after convolution and down sampling to generate boundary heatmaps. To use the geometrical and semantical information of facial shape constraint, the information transfer layers are introduced after each stack SAT. Following LAB $^{16}$, two kinds of message passing layers are utilized for information transfer: intra-level layers and inter-level layers.

**Intra-level message passing layer.** It plays a role at the end of each stack to pass information between visible boundaries and occluded ones. In the process of occlusion, the prediction of the occlusion part can be improved through the visible boundary information.

**Inter-level message passing layer.** Since different stacks get different aspects of facial information, in the case of multiple stacks, the face information is transferred in different stacks by the way of passing message from former stacks to latter stacks.

Here the implementation of message passing layers refers to the work of Chu et al. $^{29}$ As shown in Figure 2, the facial boundary is divided into $K$ branches, each represents a feature map. The calculation of boundary message passing is less than that of landmark heatmap $^{22,23}$ for the small and constant number $K$ of them.

Boundary-aware landmark localization
In order to obtain robust face alignment performance, the boundary heatmaps are fusion with the input facial image for feature learning. Here we use the stacked dense U-Nets $^9$ as the backbone for landmark localization, as shown in Figure 1(b).

**Scale aggregation topology.** The topology design for heatmap regression can obtain local and global features at different
scales, while maintaining the resolution information. Hence the stacked dense U-Nets network can get better effect by improving the topology design. As shown in Figure 3(a) and (b), the U-Net and hourglass network as classic topology design have four pooling layers. At each down-sampling step, high resolution features are acquired, then these features are combined into corresponding up-sampling features. These spatial informations are saved at each resolution through the cross layer connection. The U-Net and hourglass network achieves compelling
accuracy with a bottom-up, top-down design which endows the network with capabilities of obtaining multiscale information. Based on the U-Net and hourglass network, deep layer aggregation (DLA)\(^3\) (as shown in Figure 3(c)) augments shallow lateral connections with deeper aggregations to better fuse information across layers. Inspired by DLA, the SAT can obtain more information. As shown in Figure 3(d), SAT also adopts a bottom-up, top-down design, but there are only three pooling layers. Reducing the pool layer can greatly reduce the computational complexity and model size. In addition, the SAT adds the lower sampling input to the aggregation nodes to retain more scale feature information. Depth-wise separable convolutions\(^3\) and lateral connections\(^3\) are used to consolidate multiscale feature representations and to fuse information across layers. The SAT greatly improves the capacity of the model while the computational complexity and model size are similar to the hourglass network. Because of the performance of SAT is well in feature extraction, we use stacked dense U-Nets to get the features of images at different scales.

**Deformable convolution.** In order to further improve the capacity of the model, two U-Nets are stacked end-to-end. However, the stacked U-Nets still lacks the ability of transformation modeling due to the fixed geometry structures. Here we use the Deformable ConvNets V2\(^3\) to obtain the spatially invariant ability to the arbitrary input facial images. As shown in Figure 4, the Deformable ConvNets V2 is composed of two parts, the ordinary convolution and the migration layer. We obtain the offset of deformable convolution by convoluting the input feature map with ordinary convolution layer, and then the convolution kernel is offset by the offset field to get the deformable convolution result. The offset of Deformable ConvNets V2 can be learned in the target task without the design of features manually. In addition, Deformable ConvNets V2 adds the modulation to the offset and allocates different weights to the region modified by the offset. Accordingly, the extracted features are more concentrated in the effective region. Deformable ConvNets V2 can modify the shape of revolution kernel by learning local and dense extra offsets, thus it can have better performance on the deformation of objects, size changes and other issues.

**Image fusion with boundary heatmaps.** In order to improve the robustness in the case of occlusion and large pose, we need to fuse the boundary heatmaps with the input facial image. To fuse boundary heatmap \(H\) with an input image \(I\), the fusion function \(F\) is defined as

\[
F = I \odot (H_1 \odot I) \oplus \cdots \oplus (H_T \odot I)
\]  

(1)

where \(\odot\) is the channel-wise concatenation and \(\oplus\) is the element-wise dot product operation. The fusion operation fuses the estimated boundary heatmaps information with the original information and only focuses on the boundary information. Therefore, texture-less face regions are ignored, which can make the fusion information more effective. After fusing the information, the original input is added to keep the other valuable information.

**Adversarial learning for boundary effectiveness**

Bad boundary heatmaps will greatly deprecate the performance of the landmark localization. During training, in order to ensure the effectiveness of the heatmaps obtained in the boundary heatmaps estimation stage, we take use of adversarial learning between the estimated boundary heatmaps and the ground-truth heatmaps. For the boundary heatmaps \(H\) obtained in the boundary heatmap estimator, the generated coordinate set is defined as \(\hat{S}\), and the mapping between the coordinate and the ground-truth distance matrix is expressed as \(Dist\). Follow LAB\(^1\), we define the ground \(D_{\text{fake}}\) truth value of discriminator \(D\) for boundary heatmaps as

\[
D_{\text{fake}}(\hat{H}, \hat{S}) = \begin{cases} 
0, & \Pr(Dist(s) < \theta) < \delta \\
1, & \text{otherwise}
\end{cases}
\]  

(2)

where \(\theta\) is the distance threshold to ground-truth boundary and \(\delta\) is the probability threshold. This discriminator predicts whether boundary information is valid.

According to the idea of confrontation learning, the boundary effectiveness discriminator \(D\) and the boundary heatmap estimator \(G\) are matched. The loss of \(D\) can be expressed as

\[
\ell = -(E[\log D(M)] + E[\log (1 - D(G(I)) - D_{\text{fake}})])
\]  

(3)

where \(M\) is the ground-truth boundary heatmap. High-confidence maps can be obtained by discriminator learning, which will benefit the learning of localization network.

The training process of our method is summarized Algorithm 1.
Experiments

To validate the performance of the proposed method, extensive experiments are conducted on three popular facial image in-the-wild datasets: WFLW, 300-W and COFW.

**Datasets**

**WFLW**\(^{16}\) dataset is a challenging one, which contains 7500 faces for training and 2500 faces for testing based on WIDER Face with 98 manually annotated landmarks.\(^{16}\) The dataset is partitioned into six subsets according to challenging attribute annotation of large pose, expression, illumination, makeup, occlusion, and blur.

**300-W**\(^{17}\) provides multiple face datasets including LFPW, AFW, HELEN, XM2VTS, and IBUG with 68 automatically annotated landmarks. We use 3148 images as training samples and 689 images as testing samples. The testing images include two subsets, where 554 test samples from LFPW and HELEN form the common subset and 135 images from IBUG constitute the challenging subset.

**COFW**\(^{18}\) dataset contains 1345 faces for training and 507 faces for testing. All training face are occlusion-free while testing face are occluded partially. Each COFW face originally has 29 manually annotated landmarks.

**Evaluation metric.** To evaluate our method, we use standard normalized landmarks mean error as the evaluation criteria. We also report two further statistics: the area under the curve (AUC), and failure rate for the maximum error of 0.1. For these datasets the results are normalized by inter-pupil (eye-center-distance) distance.

**Implementation details.** All training images are cropped and resized to 256 × 256 according to provided bounding boxes. The boundary estimator is stacked four times and the dense U-Nets in localization is stacked two times. For the whole network, we set the initial learning rate to \(2 \times 10^{-5}\), and the batch size is 4, the weight decay is \(4 \times 10^{-4}\). In this process, we reduced the learning rate to \(1 \times 10^{-6}\) after 500 epochs and trained 900 epochs in total. All our models are trained with Pytorch on one TitanX GPU. Since the source code of LAB is not publicly available, here we re-implement it. Limited to the experimental environment and hyper-parameter details, we should declare that the experimental results of LAB in the following chart are not the reported results of the original work. However, the experimental environment is the same for the re-implemented LAB* and our proposed method (IBFA for short) for fair comparison, and the results of other comparable methods are reported by the original work.

**Evaluation on WFLW**

For comprehensively evaluating the robustness of our method, we report on six typical subsets of WFLW. We compared our method with the backbone LAB* and five other popular methods. The comparison results are shown in Table 1. Note that the LAB* is reported by our re-implemented work, but others are publicly reported in LAB. Though limited by the conditions, we cannot reach the reported result in LAB, but compared with other five methods, the LAB* and the proposed method IBFA still obtain the best results. The proposed method achieves 6.23% and 7.19% mean error on large expression subsets and occlusion subsets which has obvious advantages over the original method under the same conditions. In addition, the evaluation of IBFA in failure rate and AUC is better than other methods. The failure rate on large expression subsets and occlusion subsets are only 8.13% and 14.90%. The AUC on large expression and occlusion subsets are also reached 0.4188 and 0.3942 which are higher than the AUC of LAB*. This reflects the effectiveness of the improved method for exaggerated expressions and heavy occlusion as shown in Figure 5.

As shown in Table 2, We compare the two method with error rate of the five facial parts on WFLW and speed of the models. Our method performs better than LAB* in all five facial parts, especially performs well in the mouth and nose. It can be seen from Table 2 that with the increase of the scale of model, the speed of model will decrease. How to speed up the model is a problem to be solved in future.
Evaluation on 300-W

We compare our method with the backbone LAB* and eight other popular methods on 300-W dataset in the same environment. The results are shown in Table 3. Our method performs best among all of previous methods. Also, we

Table 1. Evaluation on testset and six typical subsets of WFLW (98 landmarks).

| Metric          | Method   | Testset | Pose | Expression | Illumination | Make-up | Occlusion | Blur |
|-----------------|----------|---------|------|------------|--------------|---------|-----------|------|
| Mean error (%)  | ESR      | 11.13   | 25.88| 11.47      | 10.49        | 11.05   | 13.75     | 12.20|
|                 | SDM      | 10.29   | 24.10| 11.45      | 9.32         | 13.03   | 13.75     | 11.28|
|                 | CFSS     | 9.07    | 21.36| 10.09      | 8.30         | 8.74    | 11.76     | 9.96 |
|                 | DVLN     | 6.08    | 11.54| 6.78       | 5.73         | 5.98    | 7.33      | 6.88 |
|                 | LAB*     | 6.24    | 11.42| 7.68       | 5.86         | 5.81    | 7.30      | 6.76 |
|                 | IBFA     | 5.89    | 11.36| 6.23       | 5.89         | 5.73    | 7.19      | 6.71 |
| Failure rate (%)| ESR      | 35.34   | 90.18| 42.04      | 30.80        | 38.84   | 47.28     | 41.40|
|                 | SDM      | 29.40   | 84.36| 33.44      | 26.22        | 27.67   | 41.85     | 35.32|
|                 | CFSS     | 20.56   | 66.26| 23.25      | 17.34        | 21.84   | 32.88     | 23.67|
|                 | DVLN     | 10.84   | 46.93| 11.15      | 7.31         | 11.65   | 16.30     | 13.71|
|                 | LAB*     | 11.10   | 47.13| 13.73      | 7.74         | 10.19   | 17.92     | 13.32|
|                 | IBFA     | 7.92    | 34.78| 8.13       | 5.36         | 6.80    | 14.90     | 8.94 |
| AUC             | ESR      | 0.2774  | 0.0177| 0.1981     | 0.2953       | 0.2485  | 0.1946    | 0.2204|
|                 | SDM      | 0.3002  | 0.0226| 0.2293     | 0.3237       | 0.3125  | 0.2060    | 0.2398|
|                 | CFSS     | 0.3659  | 0.0632| 0.3157     | 0.3854       | 0.3691  | 0.2688    | 0.3037|
|                 | DVLN     | 0.4551  | 0.1474| 0.3889     | 0.4743       | 0.4494  | 0.3794    | 0.3973|
|                 | LAB*     | 0.4386  | 0.1286| 0.3189     | 0.4579       | 0.4640  | 0.3684    | 0.3846|
|                 | IBFA     | 0.4728  | 0.1734| 0.4187     | 0.4862       | 0.5038  | 0.3942    | 0.4186|

AUC: area under the curve.

Table 2. Error rate of five facial parts on WFLW and the speed of models.

| Method | Chin | Brow | Nose | Eyes | Mouth | Speed (FPS) |
|--------|------|------|------|------|-------|-------------|
| LAB*   | 8.40 | 5.52 | 4.78 | 4.20 | 5.57  | 10.15       |
| IBFA   | 8.21 | 5.32 | 4.57 | 4.14 | 5.15  | 6.11        |

Table 3. Mean error (%) on 300-W common subset, challenging subset and fullset (68 landmarks).

| Method | Common subset | Challenging subset | Fullset |
|--------|---------------|--------------------|--------|
| RCPR   | 6.18          | 17.26              | 8.35   |
| CFAN   | 5.5           | 16.78              | 7.69   |
| ESR    | 5.28          | 17                 | 7.58   |
| SDM    | 5.57          | 15.4               | 7.50   |
| LBF    | 4.95          | 11.98              | 6.32   |
| DDFA   | 6.15          | 10.59              | 7.01   |
| CFSS   | 4.73          | 9.98               | 5.76   |
| MDM    | 4.83          | 10.14              | 5.88   |
| LAB*   | 4.55          | 8.28               | 5.28   |
| IBFA   | 4.22          | 7.53               | 4.89   |

Figure 5. Boundary heatmaps and landmarks of two methods in WFLW: (a) Test results of LAB* and (b) test results of IBFA.

Evaluation on 300-W

Ye et al.
compared the localization results on five parts of face on 300-W and report the mean error results in Table 4. It can be observed that performance is improved consistently in five parts of face, and the accuracy of location in eye and chin is improved most obviously.

The results of LAB* and IBFA are trained under the same conditions which reflects that the improved method obtains significantly better performance. The IBFA performs better on the challenging subset, which indicates the robustness of our method to handle occlusions and large expression. The test image of the two methods are shown in the Figure 6. We find that they are similar in the acquisition of boundary heat map, while the improved method performs better on the effect of coordinate regression. This shows that the multilayer aggregation network topology can effectively use the boundary information, thus greatly improving the accuracy of key point prediction.

**Evaluation on COFW**

We compare our method with LAB* and other six popular methods on COFW. Table 5 exhibits the experimental results we evaluated on COFW dataset. It can be seen from the results, the mean error and failure rate of our method are smaller than those methods specific to the occlusion problem, for example, HPM and DRDA. In particular, the error rate of the improved method is reduced to 2.37%, we can conclude that our improved boundary-aware face alignment model is robust against occlusion. This can also be reflected in the test result comparison Figure 7.

**Conclusions**

In this article, we propose an improved boundary-aware method using stacked dense U-Nets for robust face alignment. By utilizing the facial boundary heatmap for landmark estimation constraint, we are able to handle facial occlusions, large shape, pose, and appearance variations. Through boundary-aware landmark localization network, we can accurately regress the facial boundary heatmap to the facial feature points. Our experimental results show that the localization effect of the boundary heatmap can be

| Method | Chin | Brow | Nose | Eyes | Mouth | Speed (FPS) |
|--------|------|------|------|------|-------|-------------|
| LAB*   | 7.28 | 5.92 | 4.06 | 3.98 | 4.58  | 10.22       |
| IBFA   | 6.63 | 5.59 | 3.84 | 3.55 | 4.30  | 6.12        |

**Table 5. Mean error (%) on COFW dataset (29 landmarks).**

| Method | Mean error (%) | Failure rate (%) |
|--------|----------------|------------------|
| RCPR14 | 8.50           | 20.00            |
| HPM41  | 7.50           | 13.00            |
| CCR42  | 7.03           | 10.9             |
| DRDA43 | 6.46           | 6.00             |
| RAR44  | 6.03           | 4.14             |
| Dense U-Nets9 | 5.55 | —                |
| LAB*  | 6.61           | 7.28             |
| IBFA  | 5.42           | 2.37             |
improved by stacked dense U-Nets. In experiments, our method can achieve leading performance in WFLW, 300-W, and COFW.

**Author contributions**

JY devised the conception and design, contributed to the coding implementation, data analysis, and drafting the article; YC contributed to the conception and design, data analysis, and article drafting; XP, HZ, DG, and YR contributed to the conception and design and article revising.

**Declaration of conflicting interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Ethical approval/Patient consent**

All analyses were based on previous published studies, thus no ethical approval and patient consent are required.

**Funding**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by Zhejiang Provincial Natural Science Foundation of China (grant nos LQ18F030014, LQ18F030013, LY19F020031, and LQ17F030004) and in part by National Natural Science Foundation of China (grant no.61871350).

**ORCID iD**

Jianghao Ye https://orcid.org/0000-0001-5430-776X

**References**

1. Toderici G, Passalis G, Zafeiriou S, et al. Bidirectional relighting for 3D-aided 2D face recognition. In: *IEEE Computer Society conference on computer vision and pattern recognition*, San Francisco, CA, USA, 13–18 June 2010, pp. 2721–2728. IEEE Computer Society.
2. Taigman Y, Yang M, Ranzato M, et al. DeepFace: closing the gap to human-level performance in face verification. In: *IEEE Computer Society conference on computer vision and pattern recognition*, Columbus, OH, USA, 23–28 June 2014, pp. 1701–1708. IEEE Computer Society.
3. Lian Z, Li Y, Tao J, et al. Expression analysis based on face regions in real-world conditions. *Int J Autom Comput* 2020; 17: 96–107.
4. Belhumeur PN, Jacobs DW, Kriegman DJ, et al. Localizing parts of faces using a consensus of exemplars. In: *IEEE transactions on pattern analysis and machine intelligence*, Colorado Springs, CO, USA, 20–25 June 2011, pp. 545–552. IEEE Computer Society.
5. Zhou F, Brandt J, and Lin Z. Exemplar-based graph matching for robust facial landmark localization. In: *IEEE international conference on computer vision*, Sydney, Australia, 1–8 December 2013, pp. 1025–1032. IEEE Computer Society.
6. Zhang J, Qiu Y, Chen J, et al. Three dimensional object segmentation based on spatial adaptive projection for solid waste. *Neuro Comput* 2018; 328: 122–134.
7. Zhang J, Gui M, Wang Q, et al. Hierarchical topic model based object association for semantic SLAM. *IEEE T Vis Comput Gr* 2019; 25(11): 3052–3062.
8. Newell A, Yang K, and Deng J. Stacked hourglass networks for human pose estimation. In: *ECCV*, Amsterdam, The Netherlands, 11–14 October 2016, pp. 483–499. Springer.
9. Guo J, Deng J, Xue N, et al. Stacked dense U-Nets with dual transformers for robust face alignment. In: *Engineering applications of artificial intelligence*, Newcastle, UK, 3–6 September 2018, p. 44.
10. Zafeiriou S, Trigeorgis G, Chrysos G, et al. The menpo facial landmark localisation challenge: a step towards the solution. In: *2017 IEEE Computer Society conference on computer vision and pattern recognition*. San Francisco, CA, USA, 13–18 June 2017, pp. 2721–2728. IEEE Computer Society.

![Figure 7. Boundary heatmaps and landmarks of two methods in COFW: (a) Test results of LAB* and (b) test results of IBFA.](image)
vision and pattern recognition workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017. IEEE.
11. Cootes TF, Edwards GJ, and Taylor CJ. Active appearance models. In: Proceedings of the 5th European conference on computer vision-volume II-volume II, Freiburg, Germany, 2–6 June 1998, pp. 484–498.
12. Cristinacce D and Cootes TF. Feature detection and tracking with constrained local models. In: British Machine Vision Conference 2006, Edinburgh, UK, 4–7 September 2006, pp. 929–938. DBLP.
13. Hara K and Chellappa R. Growing regression forests by classification: applications to object pose estimation. In: European conference on computer vision, Zurich, Switzerland, 6–12 September 2014, pp. 552–567. Springer.
14. Burgos-Artizzu XP, Perona P, and Dollar PL. Robust face landmark estimation under occlusion. In: International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013, pp. 1513–1520. IEEE.
15. Cao X, Wei Y, Wen F, et al. Face alignment by explicit shape regression. Int J Comput Vis 2014; 107(2): 177–190.
16. Wu W, Qian C, Yang S, et al. Look at boundary: a boundary-aware face alignment algorithm. In: IEEE international conference on computer vision and pattern recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018, pp. 2129–2138. IEEE.
17. Sagonas C, Tzimiropoulos G, Zafeiriou S, et al. 300 faces in-the-wild: the first facial landmark localization challenge. In: Proceedings of the 2013 IEEE international conference on computer vision workshops, Sydney, Australia, 2–8 December 2013. IEEE.
18. Burgos-Artizzu XP, Perona P, and Dollar PL. Robust face landmark estimation under occlusion. In: International conference on computer vision, Sydney, Australia, 1–8 December 2013, pp. 1513–1520. IEEE.
19. Cootes TF, Taylor CJ, Cooper DH, et al. Active shape models: their training and application. CVIU 1995; 61: 38–59.
20. Tzimiropoulos G and Pantic M. Gauss-Newton deformable part models for face alignment in-the-wild. In: 2014 IEEE conference on computer vision and pattern recognition (CVPR), Columbus, OH, USA, 23–28 June 2014, pp. 1851–1858.
21. Merget D, Rock M, and Rigoll G. Robust facial landmark detection via a fully-convolutional local-global context network. In: CVPR, Salt Lake City, Utah, USA, 18–23 June 2018, pp. 781–790. IEEE.
22. Yang J, Liu Q, and Zhang K. Stacked hourglass network for robust facial landmark localisation. In: 2017 IEEE Computer Society conference on computer vision and pattern recognition workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017, pp. 79–87. IEEE.
23. Deng J, Trigeorgis G, Zhou Y, et al. Joint multi-view face alignment in the Wild. IEEE Transactions on Image Processing 2019; 1. IEEE.
24. Kowalski M, Naruniec J, and Trzcinski T. Deep alignment network: a convolutional neural network for robust face alignment. In: IEEE conference on computer vision and pattern recognition workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017, pp. 88–97. IEEE.
25. Bulat A and Tzimiropoulos G. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In: International conference on computer vision, Venice, Italy, 22–29 October 2017, pp. 1021–1030. IEEE Computer Society.
26. Chu X, Yang W, Ouyang W, et al. Multi-context attention for human pose estimation. In: Computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017, pp. 5669–5678.
27. Yang W, Li S, Ouyang W, et al. Learning feature pyramids for human pose estimation. In: IEEE international conference on computer vision, 2017, pp. 1290–1299. IEEE Computer Society.
28. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. arXiv preprint, arXiv:1406.2661, 2014.
29. Chu X, Ouyang W, Li H, et al. Structured feature learning for pose estimation. In: IEEE international conference on computer vision and pattern recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018. IEEE.
30. Yu F, Wang D, Shellhammer E, et al. Deep layer aggregation. In: IEEE international conference on computer vision and pattern recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018, pp. 2403–2412. IEEE.
31. Howard AG, Zhu M, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861, 2017.
32. Ronneberger O, Fischer P, and Brox T. U-Net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention, Munich, Germany, 5–9 October 2015, pp. 234–241. Springer.
33. Zhu X, Hu H, Lin S, et al. Deformable ConvNets V2: more deformable, better results. In: IEEE international conference on computer vision and pattern recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018. IEEE.
34. Xiong X and Torre FD. Supervised descent method and its applications to face alignment. In: IEEE international conference on computer vision and pattern recognition (CVPR), Portland, OR, USA, 23–28 June 2013, pp. 532–539. IEEE.
35. Zhu S, Li C, Loy CC, et al. Face alignment by coarse-to-fine shape searching. In: IEEE international conference on computer vision and pattern recognition (CVPR), Boston, MA, 7–12 June 2015, pp. 4998–5006. IEEE.
36. Wu W and Yang S. Leveraging intra and inter-dataset variations for robust face alignment. In: IEEE international conference on computer vision and pattern recognition workshops, Honolulu, HI, USA, 21–26 July 2017, pp. 150–159. IEEE.
37. Zhang J, Shan S, Kan M, et al. Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In: European conference on computer vision, Zurich, Switzerland, 6–12 September 2014, pp. 1–16. Springer.
38. Ren S, Cao X, Wei Y, et al. Face alignment at 3000 FPS via regressing local binary features. In: IEEE international
conference on computer vision and pattern recognition (CVPR), Columbus, OH, USA, 23–28 June 2014, pp. 1685–1692. IEEE.

39. Zhu X, Lei Z, Liu X, et al. Face alignment across large poses: a 3D solution. In: IEEE international conference on computer vision and pattern recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. IEEE.

40. Trigeorgis G, Snape P, Nicolaou MA, et al. Mnemonic descent method: a recurrent process applied for end-to-end face alignment. In: IEEE International conference on computer vision and pattern recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. IEEE.

41. Ghiasi G and Fowlkes CC. Occlusion coherence: localizing occluded faces with a hierarchical deformable part model. In: IEEE international conference on computer vision and pattern recognition (CVPR), Honolulu, HI, USA, 21 July 2017 to 26 July 2017, IEEE.

42. Feng ZH, Hu G, Kittler J, et al. Cascaded collaborative regression for robust facial landmark detection trained using a mixture of synthetic and real images with dynamic weighting. IEEE T Image Process 2015; 24(11): 3425–3440.

43. Zhang J, Kan M, Shan S, et al. Occlusion-free face alignment: deep regression networks coupled with decorrupt autoencoders. In: IEEE International conference on computer vision and pattern recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. IEEE.

44. Xiao S, Feng J, Xing J, et al. Robust facial landmark detection via recurrent attentive-refinement networks. In: Proceedings, Part I 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016, pp. 57–72.