Tweaked residual convolutional network for face alignment

Wenchao Du\textsuperscript{1, a}, Ke Li\textsuperscript{1, b}, Qijun Zhao\textsuperscript{1, c}, Yi Zhang\textsuperscript{2, d} and Hu Chen\textsuperscript{1, e, *}

\textsuperscript{1}National Key Laboratory of Fundamental Science on Synthetic Vision & School of Computer Science, Sichuan University, Chengdu, China
\textsuperscript{2}School of Computer Science, Sichuan University, Chengdu, China
\textsuperscript{a} E-mail: wenchao_du@stu.scu.edu.cn, \textsuperscript{b} E-mail: keli@scu.edu.cn, \textsuperscript{c} E-mail: qijzhao@scu.edu.cn
\textsuperscript{d} E-mail: yzhang@scu.edu.cn, \textsuperscript{e} E-mail: huchen@scu.edu.cn

Abstract. We propose a novel Tweaked Residual Convolutional Network approach for face alignment with two-level convolutional networks architecture. Specifically, the first-level Tweaked Convolutional Network (TCN) module predicts the landmark quickly but accurately enough as a preliminary, by taking low-resolution version of the detected face holistically as the input. The following Residual Convolutional Networks (RCN) module progressively refines the landmark by taking as input the local patch extracted around the predicted landmark, particularly, which allows the Convolutional Neural Network (CNN) to extract local shape-indexed features to fine-tune landmark position. Extensive evaluations show that the proposed Tweaked Residual Convolutional Network approach outperforms existing methods.

1. Introduction
Face landmark detection plays a fundamental role in many computer vision tasks, such as face recognition, expression analysis, and 3D face modeling. Recently, many methods have been proposed to address this problem, with significant progress being made towards systems that work in the wild.

Regression-based approaches \cite{4} \cite{1} \cite{20} have achieved impressive results by cascading discriminative regression functions that map facial appearance to landmark coordinates directly. In this framework, deep convolutional neural networks have proven effective as a choice for feature extraction and non-linear regression modeling \cite{17} \cite{23} \cite{21}. Although these methods can achieve very reliable results in standard benchmark datasets, they still suffer from limited performance in challenging scenarios, e.g., involving large face pose variations and severe occlusions.

In this paper, we propose an architecture named Tweaked Residual Convolutional Networks, as illustrated in Fig. 1, which can further combines the advantages of traditional cascaded regression and convolutional neural networks, and get better performance than the other state-of-the-art methods.

The proposed TRCN method, a non-linear dynamical system, consists of TCN global feature extraction module and RCN refining module instead of single CNN. Specifically, the TCN feature extraction module predicts the face shape quickly by taking holistically a low-resolution version of the detected face as the input; the following RCN module refines the landmark location by taking local patches as the input, the joint local shape-indexed feature is extracted around the current landmarks (output of the previous TCN module) to fine-tune landmark position by the RCN. As illustrated in Fig. 1, our work is also motivated by
the success of [21] and [17], however, there have a great difference from them. In summary, the contribution of our work is:

- We propose a non-linear cascaded framework to model non-linear mapping from the face image to the face shape with convolutional neural network.
- We introduce the concept of Residual Convolutional Net into the new deviation learning between the current location and the ground truth.
- TRCN can initialize the face shape better and extract shape-indexed feature with CNN, thus, it is more robust to challenging scenario than other methods.

![Figure 1. Overview of our Tweaked Residual Convolutional Network (TRCN).](image)

2. Related Work

Significant progress on facial key point detection has been achieved in recent years. Remarkably, regression based methods [20] [1] [4] [4] [1] [4] [1] [4] significantly boost the generalization performance of face landmark detection, compared to the algorithm based on statistical models such as Active shape models [13] and Active appearance models [7] [13]. A regression based approach regresses landmark locations where features extracted from face images serve as regressors directly. Landmark models are learned either in an independent manner, or in a joint fashion [21] [4].

Learning cascade-like regression models show superior performance on the face alignment task [20] [17] [21] [21]. Supervised descent method [20] learns cascades of regression models based on SIFT feature. Sun et al. [17] propose to use independent Convolutional Neural Networks (CNN) to perform coarse-to-fine shape searching. Zhang et al. [23] propose a novel approach that involves incorporating auxiliary information into the fitting process. Unlike other related methods, they do not incorporate a cascade of networks but instead frame the problem as a multi-task learning problem. Wu et al. [19] use a deep belief network to train a more flexible shape model, but do not learn any convolutional features. Finally, Baoguang Shi et al. [16] propose to jointly learn a cascade of linear regressors. Although the regressors are updated jointly via back-propagation, [16] uses linear regressors and employs hand-crafted HoG features rather than learning the features directly from the images. Zhang et al. [21] also utilise a coarse-to-fine shape search using first a global and then a set of local stacked autoencoders. However, each local autoencoder is trained with extracted shape-indexed feature (i.e.,SIFT), which seems sub-optimal. As shown in the Fig. 2, which may be not accurate enough in some special conditions.
On the contrary, in this paper, we show that it is possible to obtain large improvements when, instead of utilizing hand-crafted features, optimal features for the given problem are learnt with cascaded convolutional neural networks architecture.

**Figure 2.** Facial landmark detection under the partial occlusion, expression and pose scenario (from MTFL test set): Results of CFAN (top row) and our TRCN (bottom row).

3. **Tweaked Residual Convolutional Networks**

In this section, we present a novel Tweaked Residual Convolutional Network (TRCN) method for real-time facial landmark detection. Firstly, we introduce the overview of the proposed framework. Secondly, we describe the details about two components of TRCN, Tweaked Convolutional Network (TCN) module and Residual Convolutional Net (RCN) module. Finally we will give a detailed discussion about the difference from some existing works.

3.1. **Method Overview**

As shown in Fig. 1, the proposed TRCN designed the general regression framework by a two-level architecture, with the regression in each module modeled as a nonlinear deep network. Specifically, each module characterizes the nonlinear mapping from face image to face shape based on the shape predicted from the previous module. It is possible to obtain better improvements when optimal features for the given problem are learnt via deep convolutional network.

The first TCN module endeavors to approximate the facial landmark locations roughly, therefore, a low-resolution image is exploited for a large search step. The large step can alleviate the suffering from local minima and promise a fast model. Moreover, rather than local minima feature from mean shape, the global image feature is employed as input to avoid the inaccuracy of mean shape. In order to improve the accuracy and robustness of network prediction, we redesign global feature extraction module.

General convolutional network use a large convolutional kernel (e.g., 5×5, 7×7,) to extract the feature of the image at the initial layer of the network, and capture higher abstract feature as deeper layer, their spatial concentration is expected to decrease suggesting that the ratio of 5×5 and 3×3 convolutional kernel should increase as we move to higher layers. However, it will lose more detail feature of image when using large convolutional kernel size because of a low resolution input image, and decrease the accuracy of predicted landmark. As the result, we redesign the convolutional network with smaller kernel size (e.g., 3×3, 2×2, etc) and adjust the depth of the network, Tweaked Convolutional Networks approach ground truth of facial landmarks more accurately and robustly.

After getting an estimation face shape $S_0$ from the TCN module, the RCN makes an effort to refine the shape by regressing the residual error $\Delta S$ between the current locations and the ground truth locations step by step. The nonlinear regression model RCN is exploited to model the nonlinearity between the current feature and the ground truth shape. Meanwhile, to characterize fine variations, generally speaking, the shape-indexed feature extracted from current local shape at different scale is exploited to enforce different search region. Furthermore, the feature-extracted of all facial points are concatenated together to enforce all facial points updated jointly so as to insure a reasonable solution, even under partial occlusion scenario or large pose.
3.2. Global TCN

The TCN module, the first module of the proposed two-level framework, directly estimates the face shape based on global raw features at a low-resolution image. Given a face image $I \in \mathbb{R}^d$ of $d$ pixels, $S_0 \in \mathbb{R}^p$ denotes the ground truth locations of $p$ landmarks. The face landmark detection is to learn a mapping function $T$ from the image to the face shape as follows:

$$T : S \leftarrow I$$

Generally, $T$ is complex and nonlinear. To address this problem, $k$ single convolutional layers are stacked as a deep neural network to map the image to the corresponding shape. Consequently, the face alignment task is formulated as minimizing the follow objective:

$$T^* = \arg \min_T \| S_g(I) - t_k(\ldots t_1(I)) \|^2$$

$$t_i(a_{i-1}) = \sigma(W_i a_{i-1} + b_i) \triangleq a_i, i = 1, \ldots, k - 1,$$

$$t_k(a_{k-1}) = W_k a_{k-1} + b_k = S_0.$$  

(2)

(3)

(4)

Where $T = \{t_1, t_2, \ldots, t_k\}$, $t_i$ is the mapping function of $i$th layer in the deep network, $\sigma$ is a PReLU function and $a_i$ is the feature representations of each layer. Nonlinear mapping in term of PReLU function is employed by the first $k-1$ layers to characterize the nonlinearity between the image feature and the face shape. Therefore, linear regression is exploited in the last layer $t_k$ to get an accurate shape estimation $S_0$.

To prevent over-fitting, a regularization term $\sum_i \|W_i\|^2$ (a weights decay term) is added which tend to decrease the magnitude of the weights. The objective function is further re-formulated as bellow

$$T^* = \arg \min_T \| S_g(I) - t_k(\ldots t_1(I)) \|^2 + \alpha \sum_i \|W_i\|^2.$$  

(5)

The function $T$ contains lots of parameters and it is easy to fall into local minimum during optimization. To achieve a better optimization, firstly, we adopt the pre-train process to initialize the TCN module with part of train data, and then fine tune it with all data. As a result, the first few layers of convolutional network tend to capture the low-level feature such as texture patterns in an image, while the higher layers tend to capture abstract features containing context information of texture patterns.

After the optimization, the prediction of the facial landmarks is achieved as $S_0$, which is a rough but robust and fast approximation of the ground truth.

3.3. RCN Module

The global TCN described above gives a rough shape estimation $S_0$ of input image $I$, which is already close to the ground truth locations but not close enough due to the highly complicated variations in expression, pose, identity, etc. To achieve finer location, several RCNs are employed to iteratively predict the residual error $\Delta S_i$ between current shape $S_{i-1}$ and ground truth $S_i$ based on joint local shape-indexed features, referred as local RCNs.

Shape-indexed features extracted around the landmark points have been proven to be efficient for face alignment [4] [20] [4] [20] [20], which can only capture the information from itself, meanwhile, ignore the relevance with the other points. Most local feature is based on hand-crafted (e.g., SIFT, HOGs, LBF, etc.), as we know, it is data-driven and thus sub-optical for the face alignment task. Therefore, the facial points are modeled jointly with CNN in our RCN module, extracted-features are fused by RCN as output.

Single RCN maintains many of internal neural units that accumulates information extracted from the history of all past observations of the input space, this has the advantage that descent directions are
naturally partitioned according to the previously calculated descent directions, so it can leverage this rich information and train an end-to-end face alignment method by learning a set of data driven features with Convolutional Neural Network directly from the images in a cascaded manner. Most importantly, RCN module uses residual error to impose a constraint on the descent direction.

Similarly as the global TCN, the successive RCN is also designed as a TCN to deal with the nonlinearity of predicting the face shape, but with the extracted local patches as input. With the estimated shape $S_0$ from global TCN, the CNN feature can be extracted around each facial point, denoted as $\Phi(S_0)$. The objective of the first RCN is to achieve a nonlinear regression $R_1$ from the extracted feature $\Phi(S_0)$ to the residual error $\Delta S_1 = S_1 - S_0$ as follow:

$$R_1 = \arg \min_k \| \Delta S_1 (I) - \delta r_1 (\cdots \delta r_1 (\Phi (S_0)))) \|^2_2 + a \sum_i \| W_i \|^2_2,$$

(6)

Where $R_i = \{r'_i, r'_1, r'_2, \ldots, r'_i \}$, $\sum_i \| W_i \|^2_2$ is the weight decay term. Similar to the global TCN, the RCN is firstly initialized by using the part of train data, and then fine-tuned with all data.

After getting the face shape update ($\Delta S_1$) by the first RCN, an updated face shape can be obtained $S_1 = S_0 + \Delta S_1$, then the successive RCN extracts local features around the new shape, and optimizes a deep network to minimize the new residual error between the current location and the ground truth. The objective of the jth RCN is shown as follow:

$$R_j = \arg \min_k \| \Delta S_j (I) - \delta r_j (\cdots \delta r_j (\Phi (S_{j-1})))) \|^2_2 + a \sum_i \| W_i \|^2_2,$$

(7)

For each RCN, local features around the landmark points are extracted in a local patch of the same size but at different scale. Local patches of the different scale at low-resolution face image contain different context information and thus lead to different searching regions for each RCN. It is necessary for the anterior RCNs to approximate with a large search step when the current location is relatively far from the ground truth. On the other hand, the local patches of the same size but at smaller scale actually constrain the searching space within smaller region which means that the posterior RCN can refine the location with a tiny searching region leading to more accurate results.

4. **Discussions**

4.1. **Differences with DCNN [17] and CFAN [21]**

Both DCNN [17] and CFAN [21] follow the cascade framework and use a global nonlinear regression as the first stage to achieve a rough estimation of face shape. The differences lies on two aspects: in DCNN, after the global estimation, each facial point is refined with single CNN, which may be complicated and distort the whole shape without the constraint between facial points. While in CFAN, all facial points are refined jointly with SIFT feature, which seems suboptimal. On the contrary, the advantage of extracting image feature with CNN has been proved in the field of computer vision. Our TRCN extracts shape-indexed feature with CNN and refines jointly all facial points.

4.2. **Differences with SDM [20]**

A sequence of generic descent directions are learned by several successive RCN as well as SDM [20], but they differ in the following aspects: 1) SDM employs linear regression to model the mapping from shape-indexed features to a face shape, while our TRCN employs nonlinear regression, i.e., deep convolutional networks, to model the mapping from shape-indexed feature to face shape, which can achieve lower regression error. 2) SDM employs the mean shape as the initialization of the shape-indexed feature, which may be trapped when the initialization is far away from the ground truth, especially under the linear model. On the contrary, our TCN designs a deep convolutional network to directly predict a rough estimation of
the face shape from the global image feature rather than shape-indexed feature, and it can obtain a more accurate initialization of the shape for the following RCNs.

5. Implementation Details

5.1. Data Augmentations
To train a robust global TCN module, we augment the training data by perturbing each training sample with random changes in translation, rotation and scaling. This can effectively prevent the deep models from over-fitting and achieve robustness to various changes in the wild data.

5.2. Parameters Setting
The global TCN has seven layers with four convolutional layers and three max-pooling layers after each convolutional layer, and then followed by two linear regression layer that is capable of learning non-linear mapping from a gray-scale full face with 48 * 48 pixels to a face shape. Numbers of convolutional output in each layer are respectively 32, 32, 64, and 128. For RCNs, have five layers with three convolutional layers and two max-pooling layers after them, numbers of convolutional output in each layer of RCN are respectively 28, 48, 64 with 24*24 pixels. All training samples are pre-processed by zero mean normalization. The weight decay parameter $\alpha$ controls the relative importance of the two modules, the average sum-of-squares error term and the weight decay term. Although $\alpha$ can be set different, the same value $\alpha = 0.0005$ is used for both global TCN and RCNs for simplicity.

6. Experiments
In this section, firstly we illustrate the experimental settings for the evaluations including the datasets and contrastive methods. Secondly, we investigate the alignment results of each module in our method. Finally, we compare the proposed TRCN with the other state-of-the-art methods.

6.1. Datasets and Methods for Comparison
To evaluate the effectiveness of the proposed TRCN architecture, four public datasets are used for our experiments, i.e., MTFL [22], BioID [9], COFW [3] and LFPW [2]. BioID dataset [9] is collected under laboratory conditions, while the images in MTFL [22], COFW [3] and LFPW [2] datasets are collected in the wild environment formulating a more challenging scenario. In all experience, our TRCN is only trained with the images from MTFL training set.

We compared our framework with a few state-of-the-art methods, i.e., DRMF [1], SDM [20], DCNN [17], CFAN [21] and TCDCN [22]. The available codes from [1] predict 66 facial points, and the released code of [20] only estimates 49 landmarks located in the inner regions of the face. Both of the released codes from [17] and [21] predict 5 landmarks. [22] predicts 68 facial landmarks. Since [22] does not detect eye centers, we only compare nose tip and mouth corners with them.

All these methods were recently proposed and reported state-of-the-art performance. The normalized root mean squared error (NRMSE) is employed to measure the between the estimated facial landmark locations and the ground truth. The NRMSE is normalized by the distance between centers of eyes. The cumulative distribution function (CDF) of NRMSE is applied for performance evaluation.

| Run time (ms) | Global TCN | RCN1 | RCN2 | total |
|---------------|------------|------|------|-------|
| Run time (ms) | 3.56       | 0.94 | 0.96 | 5.46  |

6.2. Investigation of Each Module in TRCN
As the proposed TRCN method consists of two module of TCN and RCN, we investigate how each module contributes to the performance improvement for facial alignment. The experiments are conducted on MTFL
dataset [22] in terms of average detection accuracy of 5 facial points, i.e., NRMSE. The images from MTFL training set, are used for training and the test set are used for evaluation.

The evaluation results are shown in Fig. 3. As seen, the CDF of global TCN is 0.77 when NRMSE is 0.08. Therefore, if the initial estimated shape is directly used for estimating accurate facial shape, it is not accurate enough because of large pose, illustration, and so on. As the result, in the second stage, RCN module refine the landmarks accurately. In the first RCN, the detection accuracy is improved up to about 14% when NRMSE is 0.08. Little improvement when NRMSE is 0.08, and it is only up to 0.05 in the second RCN. It demonstrates that the first TCN module mainly handle the large variations due to pose and expression and the latter ones precisely refine the landmark locations in a small search region.

Figure 3. The cumulative error distribution curves of each module from MTFL

Besides the accuracy, another important factor of performance is the time complexity. We evaluate the run time of each module on MTFL testsets as shown in Table.1. The method is run on a desktop (Inter i7-4670 3.6 GHz CPU). To avoid the influence of random factors, the method is repeated several times, and the average of running time is reported.

6.3. Datasets
MTFL (Multi-Task Facial Landmark dataset) [22] has 13,466 face images, among which 5590 images are from LFW [11] and the remaining 7896 images are downloaded from the web. Each face is labeled with the positions of five key points. MTFL [22] consists of 10000 training images and 2935 test images with large variations in pose, expression, illumination, partial occlusion, etc, which makes the facial point detection quite challenging on this dataset. For our method, we only train our model with training set and evaluate each module on testset.

Bioid [9] has 1,521 images of 23 subjects, all the faces are frontal. The face detector detects faces of 1,477 images in Bioid. In this experiment, our TRCN is trained by using the images from MTFL training set. We compare with DRMF [1], SDM [20], DCNN [17] and CFAN [21] and TCDCN [22]. For a fair comparison, only the common images with face detected by all methods are employed for the testing.

The cumulative error distribution curves of these methods are shown in Fig. 4. As seen, compared to, our method reduces the detection errors significantly. Furthermore, we compare our TRCN with [21], [17] and [22] on this dataset in terms of five landmarks. As for [22], we only compare nose tip and mouth corners for a fair comparison since only the model of 68 landmarks is released. As seen, our TRCN outperforms [21] and achieve the same effect with [17].
LFPW [2] contains 1,432 face images from internet. Since some URLs of images are no longer valid, we adopt about 949 valid images for testing. All the testing images are out of our training sets. For DRMF [1] method, the tree-based face detector is used to achieve more accurate face detection. The performance of all methods are shown in Fig. 5. As seen, SDM [20] performs the best among the traditional methods. Similarly, we compare our TRCN with DCNN and CFAN in terms of five landmarks. Our approach again performs better than most methods.

COFW (Caltech Occluded Faces in the Wild) database [3] has 1,007 images of faces obtained from a variety of sources. All images were hand annotated by with 29 landmarks. It is designed to present faces in severe occlusions due to pose, sunglasses, and interaction with objects (e.g., food, hands, and hat). The performance of all methods are shown in Fig. 6. As seen, our method shows better performance.

Fig. 7 shows the results of TRCN on some extremely challenging example faces from BioID, LFPW and COFW. It shows that our algorithm is robust to the variations from pose, expression, sunglasses and partial occlusion.
7. Conclusions and Future Work
To deal with the nonlinearity in inferring face shapes from face images, we proposed the Tweaked Residual Convolutional Net with a two-level architecture, each of which figures out part of the nonlinearity. The first TCN module takes a low-resolution version of the detected face as input directly, to estimate a roughly accurate shape globally. Then, the subsequent RCNs take the local patch at different scale as input to refine the shape better. Such a strategy achieves better results than the state-of-the-art methods, such as [17] and [17], on three databases with extensive variations. Furthermore, our method can work rather efficiently, with 70+ fps even with C++ codes on a common desktop with no parallel programming.

The proposed method provides a general framework, which can be further applied to other localization-sensitive tasks, such as pose estimation, object detection, etc. In the future, we plan to further exploit the proposed Tweaked Residual Convolutional Network for broader impacts.

8. Acknowledgment
This work is supported by the National Natural Science Foundation of China (No.61202160) and the National Key Scientific Instrument and Equipment Development Projects of China (No. 2013YQ49087904).

9. References
[1] Asthana, A., Zafeiriou, S., Cheng, S., & Pantic, M. (2013). Robust Discriminative Response Map Fitting with Constrained Local Models. Computer Vision and Pattern Recognition (Vol.9, pp.3444-3451). IEEE.
[2] Belhumeur, P. N., Jacobs, D. W., Kriegman, D., & Kumar, N. (2011). Localizing parts of faces using a consensus of exemplars. IEEE Conference on Computer Vision and Pattern Recognition (Vol.35, pp.545-552). IEEE Computer Society.

[3] Burgos-Artizzu, X. P., Perona, P., & Dollar, P. (2013). Robust Face Landmark Estimation under Occlusion. IEEE International Conference on Computer Vision (pp.1513-1520). IEEE Computer Society.

[4] Cao, X., Wei, Y., Wen, F., & Sun, J. (2012). Face alignment by Explicit Shape Regression. IEEE Conference on Computer Vision and Pattern Recognition (Vol.107, pp.2887-2894). IEEE Computer Society.

[5] Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. IEEE Transactions on Pattern Analysis & Machine Intelligence, 23(6), 681-685.

[6] Cristinacce, D., & Cootes, T. F. (2006). Feature Detection and Tracking with Constrained Local Models. British Machine Vision Conference 2006, Edinburgh, Uk, September (Vol.41, pp.929-938). DBLP.

[7] Gao, X., Su, Y., Li, X., & Tao, D. (2010). A review of active appearance models. IEEE TRANSACTIONS ON SYSTEMS MAN AND CYBERNETICS PART C-APPLICATIONS AND REVIEWS, 40(2), 145-158.

[8] Gu, L., & Kanade, T. (2008). A Generative Shape Regularization Model for Robust Face Alignment. European Conference on Computer Vision (Vol.5302, pp.413-426). Springer-Verlag.

[9] Jesorsky, O., Kirchberg, K. J., & Frischholz, R. W. (2001). Robust Face Detection Using the Hausdorff Distance. International Conference on Audio- and Video-Based Biometric Person Authentication (Vol.2091, pp.90-95). Springer-Verlag.

[10] Le, V., Brandt, J., Lin, Z., Bourdev, L., & Huang, T. S. (2012). Interactive Facial Feature Localization. European Conference on Computer Vision (Vol.7574, pp.679-692). Springer-Verlag.

[11] Learned-Miller, G.B.H.E.: Labeled faces in the wild: Updates and new reporting procedures. Tech. Rep. UM-CS-2014-003, University of Massachusetts, Amherst (2014)

[12] Liu, X. (2008). Discriminative face alignment. IEEE Transactions on Pattern Analysis & Machine Intelligence, 31(11), 1941.

[13] Milborrow, S., & Nicolls, F. (2008). Locating Facial Features with an Extended Active Shape Model. Computer Vision – ECCV 2008. DBLP.

[14] Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013). A Semi-automatic Methodology for Facial Landmark Annotation. IEEE Conference on Computer Vision and Pattern Recognition Workshops (Vol.13, pp.896-903). IEEE.

[15] Saragih, J. M., Lucey, S., & Cohn, J. F. (2009). Face alignment through subspace constrained mean-shifts. IEEE, International Conference on Computer Vision (Vol.30, pp.1034-1041). IEEE.

[16] Shi, B., Bai, X., Liu, W., & Wang, J. (2014). Deep regression for face alignment. Computer Science.

[17] Sun, Y., Wang, X., & Tang, X. (2013). Deep Convolutional Network Cascade for Facial Point Detection. IEEE Conference on Computer Vision and Pattern Recognition (Vol.9, pp.3476-3483). IEEE Computer Society.

[18] Toshev, A., & Szegedy, C. (2014). DeepPose: Human Pose Estimation via Deep Neural Networks. Computer Vision and Pattern Recognition (pp.1653-1660). IEEE.

[19] Wu, Y., Wang, Z., & Ji, Q. (2013). Facial Feature Tracking Under Varying Facial Expressions and Face Poses Based on Restricted Boltzmann Machines. Computer Vision and Pattern Recognition (Vol.9, pp.3452-3459). IEEE.

[20] Xiong, X., & Torre, F. D. L. (2013). Supervised Descent Method and Its Applications to Face Alignment. Computer Vision and Pattern Recognition (Vol.9, pp.532-539). IEEE.

[21] Zhang, J., Shan, S., Kan, M., & Chen, X. (2014). Coarse-to-Fine Auto-Encoder Networks (CFAN) for Real-Time Face Alignment. Computer Vision – ECCV 2014. Springer International Publishing.
[22] Zhang, Z., Luo, P., Chen, C. L., & Tang, X. (2014). Facial Landmark Detection by Deep Multi-task Learning. European Conference on Computer Vision (Vol.8694, pp.94-108).

[23] Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2015). Learning deep representation for face alignment with auxiliary attributes. IEEE Transactions on Pattern Analysis & Machine Intelligence, 38(5), 918-930