A Deep Regression Architecture with Two-Stage Re-initialization for High Performance Facial Landmark Detection

Jiangjing Lv[1,2] *, Xiaohu Shao[1,2] *, Junliang Xing[3], Cheng Cheng[1], Xi Zhou[1]
[1] Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
[3] Institute of Automation, Chinese Academy of Sciences
{lvjiangjing,shaoxiaohu,chengcheng,zhouxi}@cigit.ac.cn jlxing@nlpr.ia.ac.cn

Abstract

Regression based facial landmark detection methods usually learns a series of regression functions to update the landmark positions from an initial estimation. Most of existing approaches focus on learning effective mapping functions with robust image features to improve performance. The approach to dealing with the initialization issue, however, receives relatively fewer attentions. In this paper, we present a deep regression architecture with two-stage re-initialization to explicitly deal with the initialization problem. At the global stage, given an image with a rough face detection result, the full face region is firstly re-initialized by a supervised spatial transformer network to a canonical shape state and then trained to regress a coarse landmark estimation. At the local stage, different face parts are further separately re-initialized to their own canonical shape states, followed by another regression subnetwork to get the final estimation. Our proposed deep architecture is trained from end to end and obtains promising results using different kinds of unstable initialization. It also achieves superior performances over many competing algorithms.

1. Introduction

Facial landmark detection, or face alignment, is to locate some predefined landmarks on the face given the face detection result, providing a representation of the face shape. It is one of the most important tasks in the field of computer vision and has been a key component of many other computer vision tasks, *e.g.*, 3D face reconstruction [1, 11], face animation [4] and face recognition [12, 3, 40].

In the past decades, computer vision researchers have devoted great efforts to solving this task, and have made significant progress [7, 8, 9, 41, 30, 33, 5, 21, 16, 29, 2, 35, 38, 39]. Among all these developments in years of studies, regression based algorithms [30, 9, 33, 5, 21, 16, 29, 2, 35, 38, 39] currently dominant the approach to solving this task. Compared with the methods using parameterized models to describe the face appearance and shape [7, 8, 41], regression based methods directly learn a series of mapping functions, *i.e.*, regressors, to progressively update the estimations of the landmark positions towards the true locations. Summarizing the results from previous studies, the pose-indexed robust features [9], the cascade regression structure [5], and the regression model [30, 21, 16], are the three most important aspects in designing a high performance landmark detection algorithm. By deploying these study results from conventional methods to the powerful deep learning framework, many promising deep learning based face alignment algorithms have been developed [35, 24, 23, 36, 27, 20].

Although great progresses have been made in the last decade, facial landmark detection still remains a very challenging problem. When the face images appear with large view variations, different expressions, and partial occlusions, even state-of-the-art algorithms may fail to locate the landmarks correctly, which restricts the applications of many facial landmark detection algorithms into practical systems. To deal with these problems, many previous work [41, 33, 27] devote much effort to learning robust image features and effective regression functions. The approach to initializing the regression based methods, however, receives relatively fewer attentions, which we believe is also crucial to the solving of this problem.

Currently most of the facial landmark detection algorithms depend on the face detection to provide a good rectangular face region as an initialization. According to recent studies [31, 32, 21], if the initial detection rectangle during testing varies from the one used in training stage, the performances of many landmark detectors degrade a lot. In many situations, users may have to choose other face detector different from the one used in training. Since different face detectors often return various face bounding boxes with

*These authors contributed equally to this study.

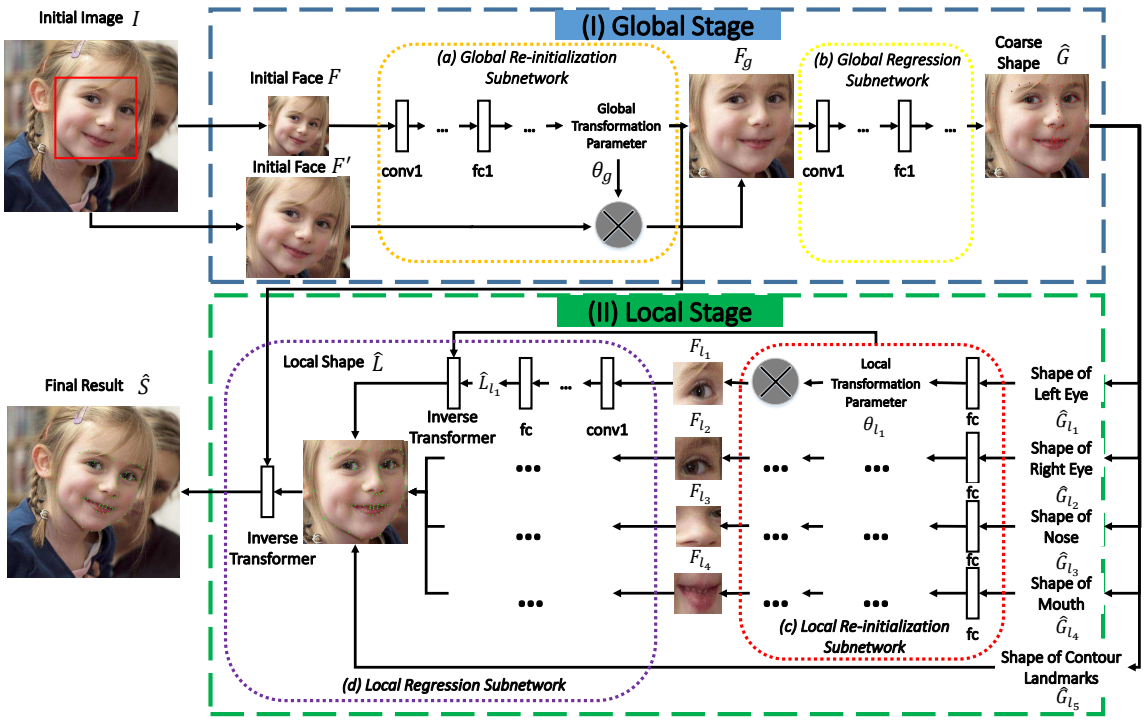


Figure 1. The pipeline of the proposed deep regression architecture with two-stage re-initialization for coarse-to-fine facial landmark detection. At the global stage (I), the face region is firstly re-initialized to a canonical shape state (a), and then regress a coarse shape (b). At the local stage (II), different face parts are further separately re-initialized to their own canonical shape states (c), followed by another regression subnetwork to get the final detection(d).

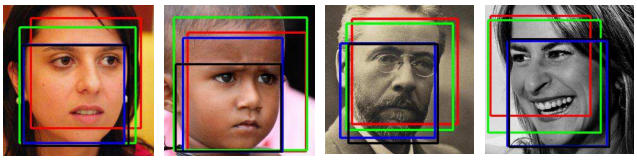


Figure 2. Different types of face bounding boxes: boxes detected by Viola-Jones detector, boxes detected by Dlib detector (green), official boxes provided by 300-W dataset (blue), boxes bounded by landmarks of ground truth (black).

different scales and center shifts (*c.f.* Figure 2), it would be very useful if a facial landmark detection algorithm can produce robust results without depending so much on the face detection results.

To explicitly deal with the initialization problem in regression based landmark detection methods, we present a deep regression architecture with two-stage re-initialization learned from end to end. Figure 1 plots the framework of the proposed deep architecture. Our two-stage re-initialization model successively re-initializes a deep regression model from coarse to fine, and global to local, to substantially boost the landmark detection performance. At the global stage, there are two subnetworks: the global re-initialization subnetwork, and the global regression subnetwork. At the re-initialization subnetwork, providing a face image with a rough bounding box, the face image is re-initialized to spatially transformed to a canonical state, *i.e.*, making the face with the same reference center, scale, and angle. This subnetwork can deal with the initialization variety problem by a large amount. With the globally normalized face state, the global regression subnetwork is trained to estimate the coarse landmark positions. There are also two subnetworks at the local stage, different parts of the face shape are further separately re-initialized to their own canonical states in the local re-initialization subnetwork, followed by the local regression subnetwork to get the final results based on the coarse landmark positions. It is very helpful to dispose the expression variations and partial face occlusions. Our whole architecture is effectively trained from end to end, exhibits more robustness to various kinds of initialization, and achieves very promising landmark detection results. On the two recent face alignment benchmarks, 300-W [22] and AFLW [17], it obtains superior results over many competing facial landmark detection algorithms.

To summarize, in this paper we make the following main contributions:

- We present a deep regression architecture with two-stage re-initialization to explicitly deal with problem of initialization variety for facial landmark detection.
- We formulate both the global and local re-initialization modules as a supervised spatial transformer learning

problem which are simultaneously trained with the whole architecture from end to end.

- We conduct extensive experiments to demonstrate that our model achieves good robustness to different kinds of initialization and state-of-the-art performances on two large benchmark datasets.

2. Related Work

Recent facial landmark detection is usually formulated as a regression problem and many of recent developments demonstrate very promising results [9, 5, 30, 21, 16, 29, 2]. With the fast development and deployment of deep learning models in computer vision, deep learning based algorithms have greatly boosted the landmark detection performance. In the following, we mainly focus on regression based and deep learning based methods, and discuss those related to our approach.

The most direct way to adopt deep learning to facial landmark detection is to let the deep model learn the features and regressors end-to-end in a cascade manner [24, 37]. To improve the performance, the architectures of these deep models are usually designed in a coarse-to-fine structure [37, 35, 23, 18, 38, 25] to progressively update the results. Some of methods jointly optimize facial landmark detection together with other tasks of facial attributes [36, 20]. These methods mainly devote efforts to learning features and regression networks, however, the initialization problem which is also important for landmark detection is ignored by these approaches.

A recent experimental study [31] evaluates and analyzes the impacts of different factors on facial landmark detection. This work shows that most methods are sensitive to different face scales, translations and initial shapes. The study in [21] finds that an "alignment friendly" detector can boost the accuracy of facial landmark detection. However, these work do not mention how to avoid bad initialization without knowing the ground truth of landmarks. We in this work are motivated to find a way which is not only robust to different poses, but also to various kinds of initialization brought by different face detectors.

The head pose assisted model [32] applies a shape which has a similar pose with real shape as the initialization. The coarse-to-fine face alignment model [13] takes a normalized full face image and then multi-scale local image patches to perform cascade regression. The competition winner model in [28] presents a progressive initialization strategy for detection which manually selects different subsets of landmarks at different regression stages. Work [38] explores the whole shape space during all the stages of the coarse-to-fine framework. The normalization and regression steps of these above mentioned work are independent to each other using a series of modules, while our network is *end-to-end* trained with *automatically learned initialization parameters*.

Recently, the Spatial Transformer Network (STN) [14] is proposed to learn instance-specific transformations of the training samples to an underlying reference sample state, which provide a way to learn invariance to different kinds of image transformations. Inspired by its good performance on the task of image classification, we present a normalization network to generate better states for the global and local facial landmark detection. Also inspired by the STN model, the DDN model [34] transforms the landmarks rather than the input image for the refinement cascade. Its point transformer network aims to learn the optimal local deformation that maps the initialized landmark to its final position, and the shape bases of the network is learned by a separated PCA procedure. Different from their work, our transformer network normalizes the input images by using the coarse landmarks and the finer landmarks learned simultaneously with the whole regression networks. Our proposed deep architecture not only learns how to provide good initialization for the global and different parts of face images, but also gets better results than that of model in [34] on benchmark datasets.

3. Our Two-Stage Re-initialization Deep Regression Model

In Figure 1, we plot the framework of our two-stage re-initialization deep regression architecture for facial landmark detection. It consists of two stages, the global stage and the local stage. In the following, we first elaborate the design of the global stage and the local stage, then introduce implementation details of the whole model.

Given an initial image I, the objective of facial landmark detection is to locate the predefined landmarks $S = [x_1, y_1, ..., x_n, y_n]^T \in \Re^{2n\times 1}$ on the face, as a registration of face shape. Different from previous work, the ground truth shapes S^* in our architecture are not fixed during training, G^* and L^* denote the target shapes at the global and local stage, respectively. In order to facilitate the subsequent formulation, we also denote the following formulation:

$$\mathbf{S} = \mathrm{M}(S) = (x_1, ..., x_n; y1, ..., y_n) \in \Re^{2\times n},$$

as the matrix form of S. In a similarly way, the matrix representations of G and L are $\mathbf{G}$, $\mathbf{L}$.

3.1. Global Stage

Previous work [21, 31] study different preprocessing steps on the face regions for facial landmark detection and find that face boxes bounding by the ground truth shapes provides the best initialization for landmark regression. Although it is unpractical to use these boxes as the initialization in real applications, our global stage, however, can take advantage of them to learn such good or even better initialization of face regions.

For an initial image I, the global stage only needs the face detector to provide a rough bounding box R to extract the face region I_R. Unlike many previous deep regression methods which directly regress the landmark locations from the face image, the global stage first learns to crop the best image region from the roughly detected face region and then learn to normalize the cropped face region to a specified canonical state with the same face size and rotation angle[1]. The cropped and normalized face image are then feed to a following regression subnetwork to get the global detection result.

3.1.1 Global Re-initialization Subnetwork

Inspired by STN, we build a global re-initialization subnetwork crops the best face region and normalizes the whole face for the following regression learning by directly learning transformation parameters θ_g. Before introducing this subnetwork, we firstly briefly review STN. It is a dynamic mechanism to spatially transform an image by producing an appropriate transformation for each input sample. It can be split into three parts: 1) a localization network, which predicts transformation parameters by taking a number of hidden layers, 2) a sampling grid, which is a set of points where an input image should be sampled to produce the transformed image, and 3) a sampler, it takes an input image and the grid to produce a transformed image.

To model the cropping and normalization operations on the input face $F \in \Re^{H\times W\times 3}$ (resized by the face image I_R to a fix-resolution image, W and H represent the width and the height of F, respectively), we employ affine transformation (other transformer functions are also feasible, *e.g.*, similarity transformation, projective transformation.) as the learning objective for the localization network. The transformation shifts the face to the center of the image, rotate the face to upright viewpoint with some skew deformations, and cut out most of unnecessary backgrounds from rough face detection result. We employ a CNN structure, *e.g.*, TCDCN[36], as the localization model to predict the transformation parameter θ_g. A transformation matrix T_{θ_g} can be constructed by the 6-dimensional affine transformation parameter θ_g, *i.e.*,

$$T_{\theta_g} = \begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{pmatrix}.$$

The transformed face image with a high resolution is denoted as image $F_g \in \Re^{H'\times W'\times 3}$, where H', W' is the width and height of F_g, larger than that of F. Then the point (x_i^t, y_i^t) on F_g can be formed by a grid using T_{θ_g} and the point (x_i^s, y_i^s) on F as follows:

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = T_{\theta_g} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}. \quad (1)$$

A bilinear sampler $\mathbf{A}(F', T_{\theta_g})$ is taken to interpolate each pixel values of F_g from the pixels around $F' \in \Re^{H'\times W'\times 3}$ which is a higher resolution image of F:

$$F_g = \mathbf{A}(F', T_{\theta_g}). \quad (2)$$

The grid generator and sampler are both differentiable, it allows gradients to be back propagated through from the sampler $\mathbf{A}(F', T_{\theta_g})$ to θ_g [14]. In the original STN model for handwriting digit recognition [14], the transformation parameters are learned from the gradients back-propagated from final classification loss. For the task of landmark detection, due to the complexity of various faces, it is difficult to guarantee the convergence of the STN model using only the following landmark regression loss. Therefore, we formulate a loss function for the transformation parameter θ as a supervised STN to speed up the convergence in the early training iterations:

$$\mathcal{L}_{\theta_g} = ||\hat{\theta}_g - \theta_g^*||_2^2, \quad (3)$$

where θ_g^* is the parameter which is able to transform F to face regions bounded by G^* (G^* represents the transformed frontal shape of the ground truth S^*). The image transformed by θ_g^* are not be able to always provide the best canonical state for the initialization. After several iterations of training when θ_g is close to the target θ_g^*, the loss function is removed to let the network continue learning $\hat{\theta}_g$ only propagated by the subsequent layers.

Examples of the canonical states are shown in Figure 3. Compared with the initial face image F, we observe that F_g has a frontal in-plane viewpoint, less unnecessary backgrounds, and it is similar to the frontal face bounded by the landmarks of ground truth, but not exactly the same.

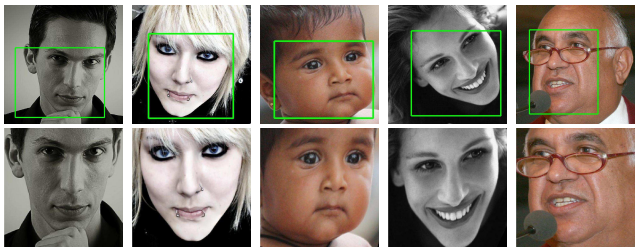


Figure 3. The results of the global re-initialization subnetwork. Top row: the input initial face images with initial face boxes. Bottome row: the transformed face images output by the global re-initialization subnetwork.

[1] In essence, any rotation angle can be used to train the model. We adopt upright face in the experiment as commonly used in other approaches.

3.1.2 Global Regression Subnetwork

With the cropped and normalized face regions, the training samples are re-initialized to a more consistent state, making the following regression learning more feasible. After F_g and G^* are obtained, a global deep regression subnetwork is introduced to learn positions of the coarse shape G in F_g. This subnetwork can be built based on a deeper structure than the previous subnetwork, *e.g.* VGG-S network [6], which comprises eight learnable layers, five among them are convolutional and the last three are fully-connected. We modify the output of last layer from 1000 to $2n$ for predicting the n landmark positions. Following the work [27], we use $L2$, normalized by inter-ocular distance instead of standard Euclidean distance as the global loss of landmark detection for faster convergence:

$$\mathcal{L}_g = \frac{||\hat{G} - G^*||_2^2}{d}, \tag{4}$$

, where $\hat{G}$ is the predicted shape in the global stage, d represents the inter-ocular distance of G^*, which is projected into the coordinate of the image F' by S^* for shape regression learning. The transformed target shape $\mathbf{G}^*$ can be obtained by using the ground truth $\mathbf{S}^*$ and inverse matrix of T_{θ_g}:

$$\mathbf{G}^* = T_{\theta_g}^{-1} \begin{pmatrix} \mathbf{S}^* \\ \mathbf{1} \end{pmatrix}. \tag{5}$$

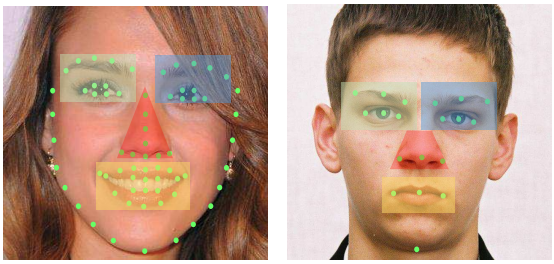
3.2. Local Stage

Due to the non-rigid property of human faces, the globally re-initialized face shape and the regressed landmark positions may still not capture all the variations of the face shape, especially for the local face parts like eyes, mouths, and noses, since they have different shapes from different identities, views and expressions. To deal with the deformations of the local face parts, we design the local stage to re-initialize the local face parts to their own canonical states for finer landmark regression, which is essential to further improve the result.

The local stage refines the shape after getting the transformed image F_g and the coarse shape $\hat{G}$. We divide landmarks of the inner face into four parts, *e.g.*, shape of left eye G_{l_1}, shape of right eye G_{l_2}, shape of nose G_{l_3}, shape of mouth G_{l_4} (as shown in Figure 4), while ignoring the part of contour landmarks for its accuracy are more difficult to improve in the local regression than that of inner landmarks. Based on the j_{th} shape part G_{l_j}, there are two subnetworks, the re-initialization subnetwork and the regression subnetwork for finer landmark detection.

3.2.1 Local Re-initialization Subnetwork

Similarly with the usage of re-initialization subnetwork in the global stage, different parts of the face shape are further separately re-initialized to their own canonical states in this re-initialization subnetwork. It consists of only one fully-connected layer, the input of which is the j_{th} shape part $\hat{G}_{l_j}$ generated by the global stage. The output of this subnetwork, which also produces a 6-dimetnsinal transformation parameter θ_{l_j} and the local transformation matrix $T_{\theta_{l_j}}$, is able to transform the face image F_g to a local normalized state F_{l_j} for the further regression of j_{th} shape part. As the canonical state for each part shown in Figure 5, most samples of the face part are aligned to the center of the patch with frontal view and retain a few of contexts.



(a) (b)

Figure 4. Four parts of landmarks in the local re-initialization subnetwork, (a) 68 landmarks of 300-W, (b) 19 landmarks of AFLW.

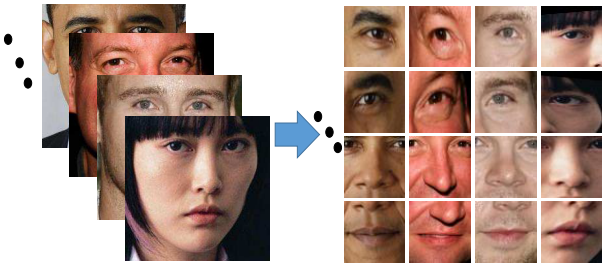


Figure 5. The results of the local re-initialization subnetwork.

3.2.2 Local Regression Subnetwork

With re-initialized face parts, the local regression subnetwork refines the shape after getting the transformed image F_{l_n} and the coarse shape $\hat{G}_{l_n}$. This subnetwork can be initialized by the first subnetwork in the global stage to minimize the loss function of shape increment:

$$\mathcal{L}_l = ||\Delta\hat{L}_{l_n} - \Delta L_{l_n}||_2^2, \tag{6}$$

where ΔL_{l_n} represents the shape increment between the two shapes: the ground truth of local shape $L_{l_n}^*$ transformed by $G_{l_n}^*$ and $T_{\theta_{l_n}}$, $\hat{G}_{l_n}^t$ transformed by $\hat{G}_{l_n}$ and $T_{\theta_{l_n}}$. The finer predicted landmarks $\hat{L}_{l_n}$ on F_{l_n} can be calculated:

$$\hat{L}_{l_n} = \hat{G}_{l_n}^t + \Delta\hat{L}_{l_n}. \tag{7}$$

For the ground truth of landmarks has a transformation while the initial face image is transformed, we add a new layer called shape inverse transformer layer, in which the

predicted shape of local stage $\hat{\mathbf{L}}_{l_n}$ on F_{l_n} can be projected to $\hat{\mathbf{S}}_{l_n}$ in the coordinate space of initial image I by using T_{θ_g}, $T_{\theta_{l_n}}$ and the rectangle geometric transformation T_R of initial face box R:

$$\hat{\mathbf{S}}_{l_n} = T_R T_{\theta_g} T_{\theta_{l_n}} \begin{pmatrix} \hat{\mathbf{L}}_{l_n} \\ \mathbf{1} \end{pmatrix}. \quad (8)$$

While omitting the variable $T_{\theta_{l_n}}$, each landmark of contour part can be projected on the initial image I from the global transformed image F_g also by using Equation 8.

3.3. Implementation Details

After the face is detected by using any kind face detector, the face bounding box is extended with a certain scale ratio of 0.2. Multiple samples are generated for each training image by disturbing the face boxes by translation and scaling, whose distributions are calculated by the differences between the initial boxes and the ground truth landmarks. As points of the sampling grids in the re-initialization networks are normalized to the range [-1, 1] by the sizes of face images, the predicted and ground truth shapes in the architecture are also transformed to the same coordinate space.

Stochastic gradient descent (SGD) is adopted for our model training. We use the min-batch size of 128, the weight decay of 0.0002, the momentum of 0.9, and iterations of 20k. There are four steps of the whole architecture training. The learning rate starts from 0.01 and 0.001 at the first three step and the last step respectively, polynomial decay is adopted for dynamically adjust the learning rate. The details of the training process are described as follows:

1. The global re-initialization subnetwork is trained at the first step. PReLU [10] is adopted as the activation functions. We warm up the training by using a small learning rate of 0.0001 at the beginning 1000 iterations and after that set it to 0.01. The input size of the subnetwork is a 60×60 resolution image.
2. Next, we fix the weights of global re-initialization subnetwork and train the global regression subnetwork. This network is initialized with an ImageNet-pre-trained model and the transformed face with size of 224×224 is employed as the input data.
3. At the third step, we fix the weights of the global stage and train the network of the local stage. The fully connected layer of each re-initialization subnetwork is initialized by using the transformation parameters which are calculated from the pre-defined canonical face parts. Each local regression subnetwork is initialized with the model of the global re-initialization subnetwork with the same input image size of 60×60.
4. At last, all the transformation parameters loss layers are removed and the whole network is fine-tuned end-to-end for shape regression.

Table 1. The comparison of NME without and with using our proposed method on 300-W dataset based on different face detectors. B_1, B_2, P^-, P indicates the results with using the $Baseline_1$, $Baseline_2$, $Propose^-$, $Propose$ method respectively.

Detectors	Common Subset	Challenging Subset	Full Set
VJ_{B_1}	8.90	14.39	9.98
Dlib_{B_1}	6.88	12.40	7.96
OD_{B_1}	5.43	8.97	6.12
GT_{B_1}	5.24	7.65	5.71
VJ_{B_2}	6.19	10.15	6.96
Dlib_{B_2}	5.30	9.13	6.05
OD_{B_2}	5.03	8.43	5.69
GT_{B_2}	5.04	7.64	5.55
VJ_{P^-}	4.95	8.36	5.62
Dlib_{P^-}	4.87	8.30	5.55
OD_{P^-}	4.56	8.16	5.27
GT_{P^-}	4.43	7.08	5.05
VJ_{P}	**4.50**	**7.89**	**5.16**
Dlib_{P}	**4.42**	**7.80**	**5.09**
OD_{P}	**4.36**	**7.56**	**4.99**
GT_{P}	**4.36**	**7.42**	**4.96**

4. Experiments

In the following sections, we first evaluate the robustness of our approach for various initialization, then compare it with other state-of-the-art methods on the benchmark datasets. In order to verify the advantages of the proposed method, we train four different models for the comparisons: the model which uses TCDCN network for training (denoted as $Baseline_1$ or B_1), the model which uses VGG-S network (see in section 3.1.1) for training (denoted as $Baseline_2$ or B_2), the model which uses the global subnetwork for training (denoted as $Proposed^-$ or P^-), and the model uses both the global and local subnetworks for training (denoted as $Proposed$ or P). The above four models are all implemented on Caffe platform [15].

4.1. Experimental Settings

In order to prove the effectiveness or our approach, we evaluate its performance on the two following benchmark datasets:

300-W [22]: The dataset consists re-annotated five existing datasets with 68 landmarks: iBug, LFPW, AFW, HELEN and XM2VTS. We follow the work [38] to use 3, 148 images for training and 689 images for testing. The testing dataset is splitted into three parts: common subset (554 images), challenging subset (135 images) and the full set (689

Table 2. The comparison of NME without and with using our proposed method on 300-W dataset based on different face extended scales (a), translations (b), rotations (c).

(a) Different Scales

Scale	0.1	0.2	0.3	0.4	0.5
$Baseline_1$	6.12	7.11	9.98	15.67	24.43
$Baseline_2$	5.69	6.27	7.05	9.59	13.54
$Proposed^-$	5.27	5.17	5.30	5.65	6.13
$Proposed$	**4.99**	**5.03**	**5.11**	**5.39**	**5.93**

(b) Different Translations

Translation	0.05	0.10	0.15	0.20	0.25
$Baseline_1$	6.29	6.96	8.51	11.86	18.67
$Baseline_2$	5.75	6.01	6.91	8.48	12.84
$Proposed^-$	5.28	5.46	5.61	5.96	6.46
$Proposed$	**5.01**	**5.15**	**5.26**	**5.36**	**5.77**

(c) Different Rotations

Rotation (°)	5	10	15	20	25
$Baseline_1$	6.35	6.98	7.91	9.35	11.71
$Baseline_2$	6.11	6.57	7.36	8.40	9.90
$Proposed^-$	5.48	5.60	5.75	6.03	6.43
$Proposed$	**5.13**	**5.24**	**5.42**	**5.77**	**6.20**

images).

AFLW [17]: It contains totally 24, 386 faces with a large variety in appearance (*e.g.*, pose, expression, ethnicity and age) and environmental conditions. This dataset provides at most 21 landmarks for each face, we ignore two landmarks of ears and evaluate our method by using the other 19 landmarks. Following the experimental settings of work [39], we use the same 20, 000 image training set and 4, 386 test set for our evaluation.

We use the normalized mean error (NME) to evaluate performance of different methods. Following work [21], inter-ocular distance is employed to normalize mean error on 300-W. As there are many profile faces with inter-ocular distance closing to zero, we use face size instead as the normalization reference on AFLW dataset.

4.2. Robustness to Various initialization

We first evaluate the impact of different face detectors for facial landmark detection on 300-W dataset. There are four types of face bounding boxes to be compared: 1) Viola-Jones (denoted as *VJ*) detector [26], a cascade face detector based on Haar-like features. 2) Dlib detector [19], an SVM detector on HOG features. Besides of them, 300-W dataset itself provides two types of face bounding boxes, 3) ground truth (denoted as *GT*), which is the tight bounding boxes of the shapes. 4) official detector (denoted as *OD*), which is very close to *GT*. Detectors of VJ and Dlib can not detect all of faces due to difficulty of some faces (*e.g.* large pose, exaggerated expression, or severe occlusion), we use the corresponding official boxes as a complement.

Table 3. The performance of our proposed method compared with other methods on 300-W dataset.

Method	Common Subset	Challenging Subset	Full Set
RCPR [2]	6.18	17.26	8.35
SDM [30]	5.57	15.40	7.52
ESR [5]	5.28	17.00	7.58
CFAN [35]	5.50	16.78	7.69
DeepReg [23]	4.51	13.80	6.31
LBF [21]	4.95	11.98	6.32
CFSS [38]	4.73	9.98	5.76
TCDCN [36]	4.80	8.60	5.54
DDN [34]	-	-	5.59
MDM [25]	4.83	10.14	5.88
$Baseline_1$	5.43	8.97	6.12
$Baseline_2$	5.03	8.43	5.69
$Proposed^-$	4.56	8.16	5.27
$Proposed$	**4.36**	**7.56**	**4.99**

The comparison of NME on 300-W dataset are shown in Table 1. It shows that *GT* provides the best initialization for landmark detection, while others have more or less decreased accuracies. It can be easily explained that the ground truth box tightly bound all the landmarks, the regression difficulty from initial images to target shapes is the smallest. Our method significantly improve the performance of the baseline under the same face detector, even under *GT*.

In order to further evaluate robustness of our architecture, we produce artificial face boxes by disturbing the official detectors with different scales, translations. We extend the face bounding boxes by a set of ratios that ranges from 0.1 to 0.5, the results are shown in Table 2 (a). Then we set a set of random ratios which are from 0.05 to 0.25 of the face box size to translate center of face box, Table 2 (b) shows the comparison of different ratios. We also rotate face images in-plane from $0°$ to $25°$ to evaluate two methods under various rotations in-plane (See Table 2 (c)). It is noted that our method is the most robust to various input with different spatial transformations.

4.3. Comparison with the State-of-the-arts

This section shows the performance of different facial landmark detection methods on the 300-W and AFLW datasets. We compare our approach with recently proposed methods [30, 33, 5, 16, 21, 2, 38, 34, 25], see in Table 3 and 4. The results show that $Proposed^-$ and $Proposed$ both get better performance than other methods on the two

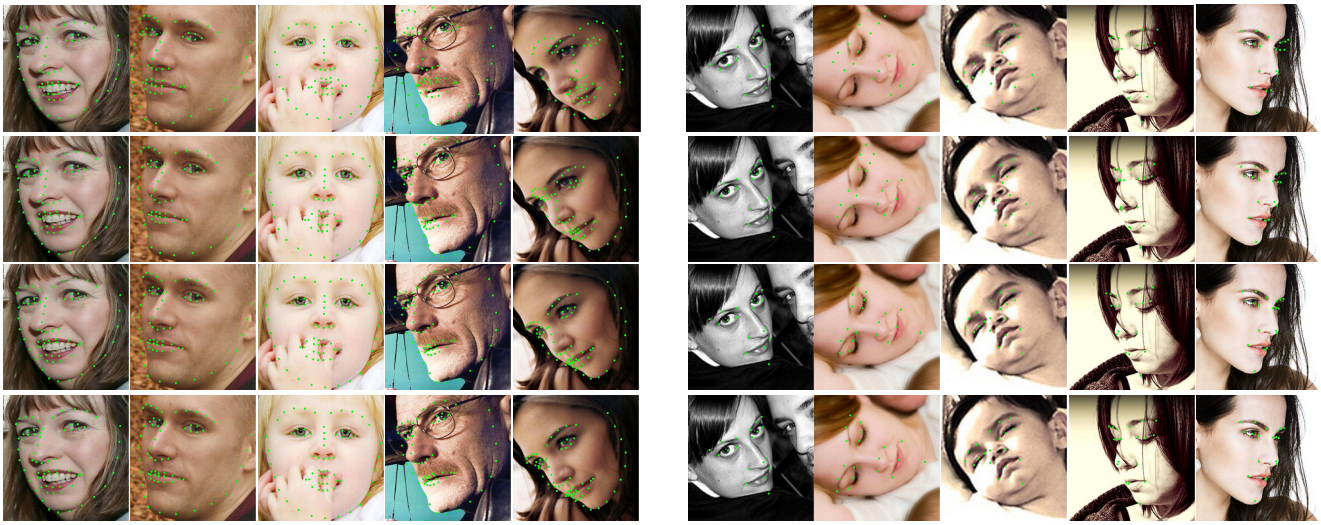


Figure 6. The comparison of facial landmark detection results on 300-W dataset (left) and AFLW dataset (right): The images are the results of $Baseline_1$ method, the $Baseline_2$ method, the $Proposed^-$ method and the $Proposed$ method from top to bottom of the single line.

Table 4. Mean Error normalized by face size on AFLW dataset compared with other state-of-the-art methods .

Method	CDM [33]	RCPR	SDM	ERT [16]	LBF	CFSS	CCL [39]	$Baseline_1$	$Baseline_2$	$Proposed^-$	$Proposed$
NME	5.43	3.73	4.05	4.35	4.25	3.92	2.72	2.99	2.68	2.33	**2.17**

datasets, and further prove that our approach is able to provide a better initialization and leads to a better landmark detection in-the-wild environment. Specifically, the fact that $Proposed$ gets the best results shows that the local stage of our method is able to further improve the accuracy of landmark detection by the re-initialization and finer regression for each shape part. The comparison of detection examples of our proposed method and the baseline methods are shown in Figure 6. The method of $Proposed^-$ and $Proposed$ are able to run at 111 FPS and 83 FPS respectively, which are evaluated based on an unoptimized Matlab interface of Caffe code with Nvidia Titan X GPU. The code and models will be made publicly available online.

5. Conclusion and Future Work

In this paper, we focus on improving the initialization part of landmark detection, which is ignored by most previous work. We present a deep regression with two-stage re-initialization architecture which is more robust to various kinds of initialization and achieves the state-of-the-art performance on benchmarks of landmark detection. From the global stage to the local stage, the initial face images are transformed to the normalized states which are more insensitive to various input derived from different face detectors and more suitable for finer landmark localization. In the future, we will continue to improve our detection performance by introducing more flexible transformations, *e.g.*, 3D transformation, and explore an end-to-end architecture to directly detect landmarks from an input image even without face detection module.

Acknowledgement

This work is partially supported by the Project from National Natural Science Foundation of China (Grant No. 61672519, 61502444, 61602433, 61472386), Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA06040103), and Chongqing Research Program of Basic Research and Frontier Technology (No. cstc2016jcyjA0011). Two Titan X GPUs donated by NVIDIA Corporation are used for this research.

References

[1] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. Rohith. Fully automatic pose-invariant face recognition via 3d pose normalization. In *IEEE International Conference on Computer Vision*, 2011.

[2] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *Proceedings of IEEE International Conference on Computer Vision*, 2013.

[3] T. Berg and P. N. Belhumeur. Tom-vs-pete classifiers and identity-preserving alignment for face verification. In *Proceedings of British Machine Vision Conference*, 2012.

[4] C. Cao, Q. Hou, and K. Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on Graphics*, 33(4):43, 2014.

[5] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014.

[6] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, abs/1405.3531, 2014.

[7] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[9] P. Dollar, P. Welinder, and P. Perona. Cascaded pose regression. *Proceedings of IEEE International Conference on Computer Vision*, 238(6):1078–1085, 2010.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of IEEE International Conference on Computer Vision*, 2015.

[11] Y. Hu, D. Jiang, S. Yan, L. Zhang, and H. Zhang. Automatic 3d reconstruction for face recognition. In *Processings of IEEE International Conference on Automatic Face and Gesture Recognition*, 2004.

[12] G. B. Huang and E. Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, 2014.

[13] Z. Huang, E. Zhou, and Z. Cao. Coarse-to-fine face alignment with multi-scale local patch regression. *arXiv preprint arXiv:1511.04901*, 2015.

[14] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, 2015.

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[16] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[17] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Processdings of IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[18] H. Lai, S. Xiao, Z. Cui, Y. Pan, C. Xu, and S. Yan. Deep cascaded regression for face alignment. *ArXiv e-prints*, 2015.

[19] A. K. McCallum. Mallet: A machine learning for language toolkit. "http://www.cs.umass.edu/ mccallum/mallet", 2002.

[20] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *arXiv preprint arXiv:1603.01249*, 2016.

[21] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment via regressing local binary features. *IEEE Transactions on Image Processing*, 25(3):1233–1245, 2016.

[22] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of IEEE International Conference on Computer Vision Workshops*, 2013.

[23] B. Shi, X. Bai, W. Liu, and J. Wang. Deep regression for face alignment. *arXiv preprint arXiv:1409.5230*, 2014.

[24] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[25] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4177–4187, 2016.

[26] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[27] Y. Wu and T. Hassner. Facial landmark detection with tweaked convolutional neural networks. *arXiv preprint arXiv:1511.04031*, 2015.

[28] S. Xiao, S. Yan, and A. A. Kassim. Facial landmark detection via progressive initialization. In *Proceedings of IEEE International Conference on Computer Vision Workshops*, 2015.

[29] J. Xing, Z. Niu, J. Huang, W. Hu, and S. Yan. Towards multi-view and partially-occluded face alignment. In *Proceedings of IEEE International Conference on Computer Vision*, 2014.

[30] X. Xiong and F. Torre. Supervised descent method and its applications to face alignment. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[31] H. Yang, X. Jia, C. C. Loy, and P. Robinson. An empirical study of recent face alignment methods. *arXiv preprint arXiv:1511.05049*, 2015.

[32] H. Yang, W. Mou, Y. Zhang, I. Patras, H. Gunes, and P. Robinson. Face alignment assisted by head pose estimation. In *Proceedings of British Machine Vision Conference*, 2015.

[33] X. Yu, J. Huang, S. Zhang, W. Yan, and D. Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *Proceedings of IEEE International Conference on Computer Vision*, 2013.

[34] X. Yu, F. Zhou, and M. Chandraker. Deep deformation network for object landmark localization. *arXiv preprint arXiv:1605.01014*, 2016.

[35] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *Proceedings of European Conference on Computer Vision*, 2014.

[36] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *Proceedings of European Conference on Computer Vision and Pattern Recognition*. 2014.

[37] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin. Extensive facial landmark localization with coarse-to-fine convolutional

network cascade. In *Proceedings of IEEE International Conference on Computer Vision Workships*, 2013.

[38] S. Zhu, C. Li, C. C. Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of IEEE International Conference on Computer Vision*, 2015.

[39] S. Zhu, C. Li, C. C. Loy, and X. Tang. Unconstrained face alignment via cascaded compositional learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[40] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[41] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012.