World Wide Web (2019) 22:1481–1498
https://doi.org/10.1007/s11280-018-0615-9

Spatial alignment network for facial landmark localization

Huifang Li[1] · Yidong Li[1] · Junliang Xing[2] · Hairong Dong[3]

Received: 8 March 2018 / Revised: 2 May 2018 / Accepted: 18 June 2018 /
Published online: 7 August 2018

Abstract Facial Landmark Localization (FLL) on unconstrained images still remains challenging as they poses complex variation in face spatial structure and appearance. To address this problem, we propose a Spatial Alignment Network (SAN), which consist of two modules, like the transformation sub-network and the estimation sub-network. In the first module, we propose two methods to achieving spatial transformation, one is the handcrafted method which can ensure model stability and the other is the learning-based method which is efficient and flexible. In the second module, we add an attention layer in the deep CNN to enhance the importance of discriminative features and obtain more accurate results. Through extensive experiments, our model achieves good performance on several public challenging datasets.

Keywords Facial landmark localization · Spatial transformation · Canonical shape · Attention · Convolution neural network

✉ Yidong Li
ydli@bjtu.edu.cn

Huifang Li
17112084@bjtu.edu.cn

Junliang Xing
jlxing@nlpr.ia.ac.cn

Hairong Dong
hrdong@bjtu.edu.cn

[1] School of Computer and Information Technology, Beijing Jiaotong University, Beijing, 100044, China

[2] Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China

[3] Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, 100044, China

1 Introduction

As one of the critical issues in face recognition, facial landmarks localization (alias face alignment) is an active area in computer vision, which obtains face shape by locating the predefined key points (e.g.,eye corners, nose tip) on human face automatically [10]. The performance of face alignment accuracy may have significant impact on many face recognition tasks, such as tracking [4], 3D face reconstruction [12], and face anti-spoong [17, 19]. For example, Li et al. [17] applies eye blinking (eye position) to detect fake faces and avoid presentation attack. Existing studies have yielded satisfactory localization results under certain constraints [5, 10, 33]. In real world scenarios, many images collected in the wild, and the detected images are suffer from spatial distortion introduced by perspective irregularities where the positions in camera with respect to the scene alter the dimensions of the scene geometry. So, spatial variation among images collected in the wild are severe which degrade most methods heavily.

To address this problem, existing studies mainly use the cascaded framework proposed in [10] to approach the ground truth progressively through multi-stages. These methods can be divided into two categories: one is based on hand-crafted feature, which may be indiscriminate and unreliable [3, 5, 10, 22]; and the other applies deep convolution neural network (CNN) [23, 29, 34] to learn high-level feature, which achieves excellent performance in some complex tasks. Most networks can extract discriminating features from various appearance and spatial information through: 1) hierarchical convolution layers to learn non-linear transformation; 2) spatial pooling layers to preserve spatial invariance; and 3) data augmentation methods. However these approaches come with several shortcomings. First of all, the cascaded structure is accurate but generally with low efficiency and high calculation cost. Moreover, convolution layers often have hundreds of channels to capture various information which may be confusing in some extent. Spatial down-pooling layers can reduce model complexity but are with limited tolerance to geometric variation. Down-pooling also destroys spatial information in images which is crucial to subsequent layers learning. In addition, data augmentation techniques try to enhance model tolerance to geometric distortion through synthesizing multiple new training samples, but it fails in fitting all real world images.

We propose a novel spatial alignment network (SAN) that eliminates the spatial and appearance variation in the picture and enhances discriminating features to accurately locate facial landmarks on the image collected in the wild. This network consists of two sub-networks, and the first one mainly implements image transformation and the other predicts the landmark position. In order to achieve image transformation, we propose two methods, and one is to manually calculate transformation parameters to obtain a stable result, and the other is to inference transformation parameters by learning-based method to increase the efficiency and the flexibility. In order to estimate the landmark position, we add an attention layer to the network to enhance the importance of discriminative features and obtain more accurate results. The main contributions of our work are as follows:

- We propose the spatial alignment framework to eliminate spatial and appearance variation in the image and resolve misalignment in deep CNN model.
- To achieve image transformation, we propose two methods to get transformation parameters, including hand-crafted method to guaranteed stability and learning-based method to improve efficiency and scalability.
- We integrate an attention layer to enhance significant feature intensities.

In addition, our model get accurate localization accuracy in some challenging dataset, and it can easily be extended to the cascade framework.

The remainder of this paper is organized as follows. Section 2 provides a brief survey of the facial landmark localization and the cascade framework. Further, we explain our proposed Spatial Alignment Network in Section 3. Experimental details and results are shown in Section 4. Finally, we draw a conclusion in Section 5.

2 Related works

In this section, we review the nominal approaches to facial landmark localization (FLL) and detection. It is a mature computer vision problem with multiple research works. The classical methods include Active Appearance Models [9], Constrained Local Models [1, 28] and Cascaded Regression Models [5, 10]. However, regression-based methods are more efficient and popular than the others as they need less prior information. Our proposed method is also under this idea. Basically, the regression is to build a mapping from the extracted features to the target label. For FLL, we learns a function in form of Eq. 1 [5]. The cascaded regression is to build stacked multi-functions, which approaches the target progressively as in Eq. 2. The symbol S is the face shape, which is a $2n \times 1$ vector consisting of n positions $(x_i, y_i), i = 1, ..., 68$, S^0 is the initial mean shape, and I_i represents the ith image. In addition, Φ and F are feature extraction function and regression function, which mapping from the image space to the feature space and transforming the feature space to the target shape respectively. In the cascaded formulation, t denotes the stage number, and current stage regression outputting depends on the previous output.

$$S = F(\Phi(S_i^0, I_i)) \tag{1}$$

$$S^t = S^{t-1} + F^t(\Phi^t(S_i^{t-1}, I_i)) \tag{2}$$

Regression-based methods can be classified into two kinds based on feature learning methods, including traditional hand-crafted features and deep features. The shallow SIFT [33], HOG and Pose/Shape-index [5] feature are efficient enough to analyze images collected in limited conditions but perform poorly confronting images in the wild. In addition, the traditional cascaded regression depends on the mean shape S^0 heavily that results in the local minimum solution. With the notable success of deep learning networks in computer vision tasks [11, 36], researchers extend deep methods into FLL field [20, 24, 29, 35]. Deep networks' parallel hierarchically structure and activation function contribute to learn multiple non-linear mapping functions which provide discriminating representation for varied images.

2.1 The cascaded network based on convolution feature

To give more explanation, we have introduced a cascade framework to solve FLL problem in the previous paper [16], as seen in Figure 1. There are 3 stages stacked sequentially, and the output in current stage is the input in the next stage, so as to refine predicted position gradually. That is, the task of the current stage is to approximate the deviation between the ground truth and the estimation in the previous stage except the first stage. In the first stage, the model's target is to directly estimate all landmarks positions from scratch based on the global convolution feature. The feature is extracted from a deep CNN, which concatenates low layers with high layers to preserve more localization information. In the last two stages,

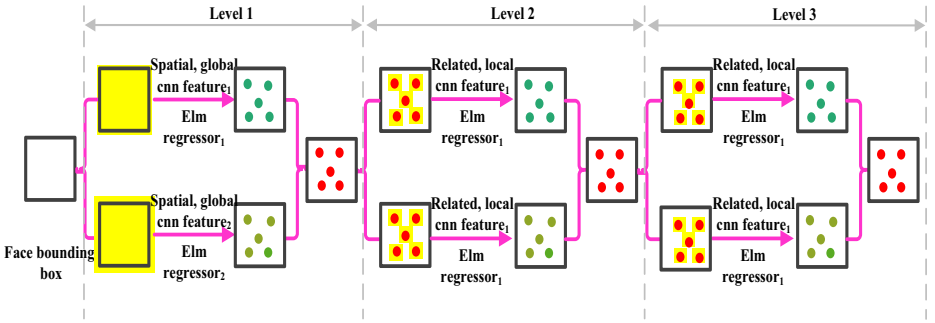

Figure 1 The three-stage framework for coarse-to-fine facial landmark localization. The input region is the face region detected by the face detector, the black square is the detected face bounding box. Yellow shaded areas are input regions of network. Red dots are the final predictions at each stage, green dots are predictions outputted by a single network. We fuse two predictions in each stage

we build a shallow convolution network for each landmark to extract local features to refine the estimation.

This work mainly studies the effect of convolution feature and regressor, allowing us to gain an insight into facial landmark estimation task, such as how to preserve localization information in deep network and how to define a model. The proposed model in the previous work can provide an accurate estimation for 5 landmarks, and it is more suitable for non-real-time tasks and simple face analysis.

To analyze more complex and real-time facial tasks, we investigate more information critical for localizing 68 facial landmarks in this paper, such as eliminating spatial and appearance variation among unconstrained faces. There are some techniques to relax and eliminate the variation such as the classical image transformation approaches and the notable Spatial Transformer Networks (STNs) [18, 21]. The classical image transformation applies a planar affine/similarity/projective transformation [31] to a distorted image, which imposing the translation, rotation, isotropic scaling and shear to images. And the STNs provides a novel strategy to integrate the transformation into the neural network and validate such solution is differential to back-propagate to a canonical image. It is provided for image/object classification firstly.

In this paper, we extend the above methods into localization tasks. In addition, for varied images, it also is important to pick most significant intensities, and this idea is similar to human attention mechanism [8], so we add an attention layer in our method.

3 Spatial alignment network

In this section, we describe our proposed Spatial Alignment Network (SAN). As difficulties of analyzing face images collected in the wild are mainly caused by the spatial and appearance variation. To address this problem, we introduce a spatial alignment network eliminating spatial variation and enhance significant feature intensities. Our SAN model mainly consists of two parts, such as the transformation sub-network for aligning images to the canonical face and the estimation sub-network for locating landmarks. Specifically we propose two methods to compute transformation parameters and convert the image. The first one manually computes these parameters based on some fixed source and target points that

ensuring model's stability, and the other method builds a learning network to inference all parameters which is efficient. In addition, we extend an attention layer into our estimation sub-network to extract more discriminating feature.

3.1 The proposed framework

The pipeline of our proposed Spatial Alignment Network(SAN) is illustrated in Figure 2. The face bounding box obtained by the face detector is fed into the transformation sub-network. This module is designed to achieve spatial transformation, which includes a CNN and a warping component, and we propose two methods to realize the transformation. After spatial transformation, the face is aligned with rotation, scale and transformation operation that eliminates in-plane variation and reduces the difficulty of analysis. Then the transformed face is fed into the estimation sub-network to localize 68 facial landmarks. In this part, the attention map is merged into our deep CNN to refine appearance representation. In the following sections, we present the two major components of our framework in details and also give a overall analysis.

3.2 The spatial transformation

Conflicts In order to learn a model which can align facial landmarks under unconstrained condition (such as various poses, illumination, expressions and occlusions), training data has to contain a lot of faces covering all possible variation. Although it is achievable to learn this model, the training needs all kinds of faces images. In addition, the learning is a quite difficult task when there are large variation among the training images, and either a complicated mapping is required, or the accuracy will be compromised. As it is common that increased model complexity results in poorer generalization performance [31]. This means that a simpler or more regularized model is favorable, which trained on a limited range of variation but align all possible poses.

In order to balance this conflict, motivated by transformation in-variance, we design a spatial transformation module to eliminate spatial variation among training samples so that the estimation model still be able to align faces in an arbitrary pose. This module is essentially a trade-off between the structural complexity and the prediction accuracy, which contains a localization network and warping component. These two components are implemented by two different methods, and each component has a different role in different methods. The first method is called handcraft transformation method, which computes transformation parameters by some fixed points. Another is named learning-based transformation method, which outputs transformation parameters by network inference.

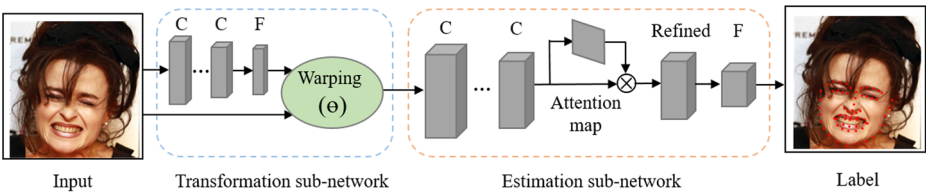

Figure 2 The general pipeline of our proposed spatial alignment network. There are two main modules, including the transformation sub-network to convert images and the estimation sub-network to predict the landmark position. The warping module in the sub-transformation network connects these two modules together, and we propose two methods to achieve the warping operation

The above mentioned transformation parameters are from affine transformation. It is an important kind of linear 2-D geometric transformation which maps a pixel intensity value located at position x^s, y^s in an input image into a new position x^t, y^t in an output image. The new pixel is computed by interpolation method. The transformation is a linear combination of translation, rotation, scaling and/or shearing (non-uniform scaling in some directions) operations. After affine transformation, some variation can be eliminated or distorted images can be corrected. The general affine transformation is commonly written in homogeneous coordinates as shown in Eq. 3. By only defining B matrix, it is pure translation operation and A is a unit matrix. Pure rotation uses the A matrix, $A = \begin{bmatrix} cos(\theta) & -sin(\theta) \\ sin(\theta) & cos(\theta) \end{bmatrix}$. Similarly, pure scaling is $A = \begin{bmatrix} a_1 & 0 \\ 0 & a_4 \end{bmatrix}$.

$$\begin{bmatrix} x^t \\ y^t \end{bmatrix} = A \begin{bmatrix} x^s \\ y^s \end{bmatrix} + B, A = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}, B = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} \tag{3}$$

As this paper proposes two methods to implement affine transformation, we give details of these two methods in following sections.

3.2.1 The handcrafted transformation

In this part, we provide a detailed description of the handcrafted transformation method, as shown in Figure 3, which calculate affine parameters manually to align the varied input image to the canonical face shape. Figure 4 is the canonical shape $\overline{S}$, which is the mean location of each landmark point in all training samples.

According to the explanation of affine transformation, at least three source points and three target points are required to compute six transformation parameters. Source points are on the original image and target points are on the canonical image. Since the test sample has no additional information besides the image, we first build a convolution network to locate eight instead of 68 key points in the face image. These eight points are outer/inter eye corner, and nose tip, and left/right mouth corner and bottom lip center. There are three reasons for this. The first is that these eight points can identify an individual image, and it is enough to calculate affine matrix. The second is efficiency, because the network that locates eight points is smaller and faster than the network of 68 points. The last is the accuracy. These points are in the inter face and have more distinguishing features. They are easy to detect, which can reduce the dependency of the subsequent steps.

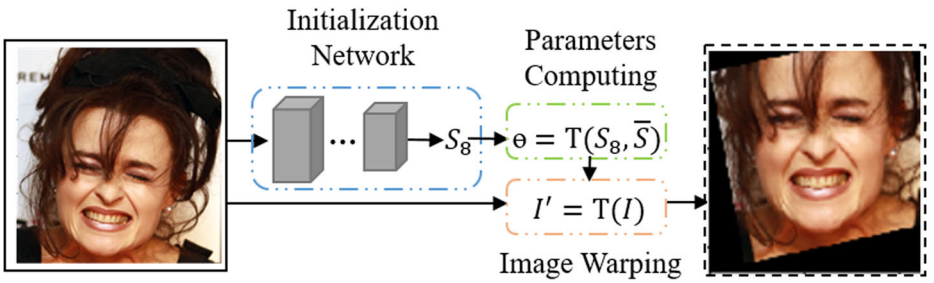

Figure 3 An illustration of the handcrafted method. It consists of three parts: **a** initialization network, **b** parameters computing, and **c** image warping

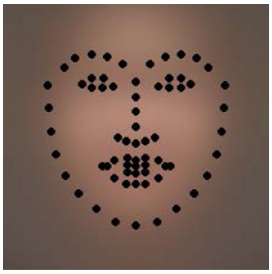

Figure 4 Canonical/Mean Face and Shape. It is an average of all training images (after uniform scaling, rotation and transformation)

After getting source and target/canonical points, we begin the affine transformation. For each image with $(x_1^s, y_1^s, ..., x_8^s, y_8^s)$ which annotated by the initialization model, we manually compute affine parameters $\theta = T(S, \overline{S})$. Next, we get the transformed image I' and shape S' based on the canonical coordinate frame by applying θ. Examples of the transformed images is shown in Figure 3. Compared with the original face image, we can see that the transformed image has a frontal viewpoint and less background information, which is learning easily.

3.2.2 The learning-based transformation

The handcrafted transformation proposed above is straightforward and transparent but not scalable and efficient because three transformation parts in the method are independent. In addition, the computing transformation parameter depends on the initialization network heavily, which can impose an adverse impact on the following estimation if the network outputs some inaccurate landmarks. So, it is better to build an end-to-end model so that the transformation and the estimation components can interact.

To achieve that, we extend the Spatial Transformer Networks (STNs) [2] to directly learn transformation parameter θ in the unlimited face alignment task, as shown in Figure 5. We fist briefly review STNs. STNs [2] consists of three parts: 1) a localization network, which inference the transformation parameters by several stacked hidden layers, 2) a sampling grid, which is multiple points where an input feature map should be sampled to generate the transformed feature map, and 3) a sampler, it takes the grid and the input feature to generate

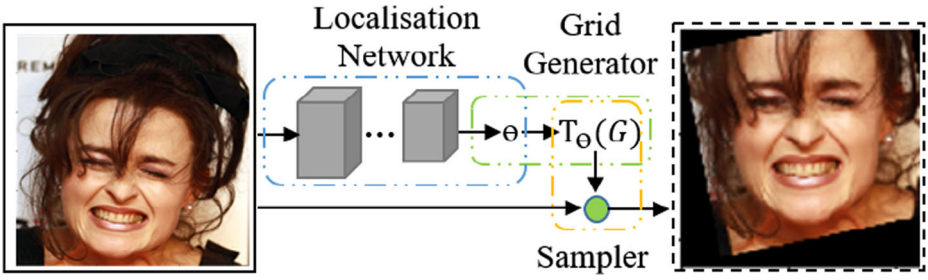

Figure 5 An illustration of the learning-based method. It consists of three parts: **a** a localization network, **b** a grid generator, and **c** a sampler

the aligned feature. Therefore, STNs can transform an image dynamically by learning an appropriate transformation parameters.

To align the varied face image, we first design a localization network and apply affine transformation to predict six affine parameters θ. Then, the point on the transformed image can be formed by a grid with θ and the point in the input image. A bilinear sampler is exploited to interpolate each pixel value in the transformed image. Examples of the transformed images is shown in Figure 5. Compared with the original face image, it is obvious that the transformed image shows a frontal in-plane viewpoint, which is favorable for analyzing.

3.3 Spatial attention

Current CNN models generally end with a global average pooling and some fully connected layers. The global average pooling operation averages all pixel values on each feature map which weakens the effect of significant pixels. To handle this problem, we exploit attention mechanism which weights the importance of significant pixels [7].

We extend a soft attention layer in our network, which can be merged into CNN and trained end-to-end [6]. To achieve soft attention, we first get the summarized feature as Eq. 4, where f denotes convolution feature maps, $*$ denotes convolution operation, W^a indicates convolution filter parameters, g is the non-linear activation function ($Sigmoid$), and $s \in R^{H \times W}$ represents the summary of all feature maps in f. Then, we normalize s using softmax operation as Eq. 5, where $s(x, y)$ is pixel value at position (x, y), $x = 1, ..., W$, $y = 1, ..., H$, the output $\psi(x, y)$ is the attention map, and $\sum_{(x,y)\in(W,H)} \psi(x, y) = 1$. Finally, the attention map is applied to each channel of feature f as Eq. 6, where index c indicates feature channel number the symbol $\star$ represents channel-wise Hadamard matrix product operation, and f^{att} represents the final refined feature, which is actually re-weighted by the attention map. The refined feature f^{att} has the same size as feature f.

$$s = g(W^a * f + b) \tag{4}$$

$$\psi(x, y) = \frac{e^{s(x,y)}}{\sum_{(x,y)\in(W,H)} e^{s(x,y)}} \tag{5}$$

$$f^{att} = \psi \star f, f^{att}(c) = f(c) \circ \psi \tag{6}$$

After refining, each pixel is re-weighted so that significant pixels approach 1 and undiscriminating pixels approach zero. In addition, the attention map can be used as feature selectors in the feed-forward inference and gradient update filters in the back-propagation. The STNs used for spatial transformation is actually a special kind of attention mechanism, which transforming or refining input image/feature by transformation parameters generated by localization network.

3.4 Network architecture

Our previous cascade work [16] proves the effectiveness of connecting the lower and upper layers of the CNN in localizing facial landmarks. This is because the hierarchical structure of CNN networks has some special properties, such as lower layers respond to edges and corners with better localization properties, and higher layers tend to learn more complex

and class-specific representation. For localization tasks, such as facial landmark localization, they require pixel-level accuracy, which need a lot of localization information, thereby making full use of the intermediate layers is necessary; CNN models exploited in this paper take this property into consideration.

Transformation sub-network In this paper, we use hyperface [23] network to achieve spatial transformation. More specifically, in the manual conversion method, it is used to regress eight facial landmarks to get the source points in the input image, and in the learning-based transformation, it is the localization network to directly inference 6 affine transformation parameters. The architecture of hyperface is shown in Figure 6, which fuses P_1, C_3, and P_5 layer of AlexNet [15] by concatenating their features. As they can't be concatenated directly with different size, the author adds C_{1a} and C_{3a} after P_1 and C_3 respectively to obtain consistent feature maps as P_5. After concatenating, there are 768 channels in the network, and its dimensions are a bit high, so a convolution layer is added to reduce dimensions to 192, its feature maps size remain unchanged. Then, there are three fully connected layers followed to regress the landmark position or inference affine parameters. In addition, each convolution layer and fully connected layer is followed by a non-linear activation function ReLu (Rectified Layer Unit), which converges faster than Sigmoid.

Estimation sub-etwork As shown in Figure 2, the estimation sub-network part outputs final 68 facial landmarks. In this paper, we apply a deep CNN named ResNeXt [32] as regressing 68 points is more difficult than eight inner face points and the five convolution layer stacked hyperface lacks capacity. The ResNeXt [32] improves the classical ResNet [11] aggregateing a set of transformation with the same topology to get a homogeneous, multi-branches architecture. This improvement makes only a few hyper-parameters to set.

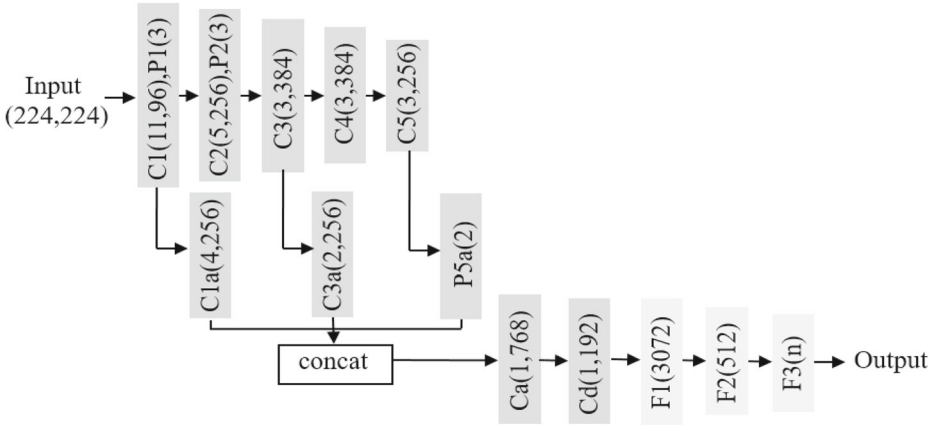

Figure 6 The structure of our baseline network hyperface [23]. It extends the classical AlexNet [15]. The depth of hyperface is five, including five convolution layers C_1, ..., C_5 followed by nonlinear activation Relu where two inside indexes indicate filter kernel size and the number of filters. P_1, P_2 and P_3 are three max pooling layers where the inside index indicates the pooling stride. In addition, P_1, C_3 and P_5 are concatenated to preserve spatial information, which applying C_{1a} and C_{3a} to change the feature map size of P_1 and C_3 in order to have the same size as P_5. Layer C_a and C_d are used for concatenating and dimension reduction separately. Finally, F_1, F_2 and F_3 are three fully connected layers, the inside index means the number of neural units

Through experimental, we finally integrate 17 convolution layers in it. In addition, experimental results of the hyperface and ResNeXt networks also verify the necessity of using the depth model in the 68 points estimation phase.

Based on the basic ResNeXt network, we add an attention layer in the basic ResNeXt network for extracting discriminating feature and promoting convergence, as shown in Figure 7, and our experimental result validates its effectiveness.

3.5 Overall analysis

Method analysis In this paper, our proposed model is a one-stage method, and it can be extended into the cascaded framework easily. Specifically, to build a three-stages model, we can directly adopt our SAN in the first stage to directly output 68 landmarks. In the second and third stage, we simply alter the target of the SAN to a deviation between the ground truth and the prediction of the previous stage. So, our model is highly scalable.

Compared with other works Some existing methods also consider spatial conversion, such as the DAN [14] model which employing similarity transformation and the MIX [31] method which computing affine transformation parameters by hand-crated feature and average shape. Compared with these methods, there are three main differences. Firstly, our paper proposes two methods to obtain the transformation parameters, then compares advantages and disadvantages of them with the theoretical and experimental analysis, while other papers only use one method. Secondly, for the hand-crafted affine transformation, to get the source position on the original image, we train a CNN to initialize the key-point directly, however, other papers train the model based the mean shape that is not robust as it depends on the mean shape heavily. Thirdly, for the learning-based transformation, we

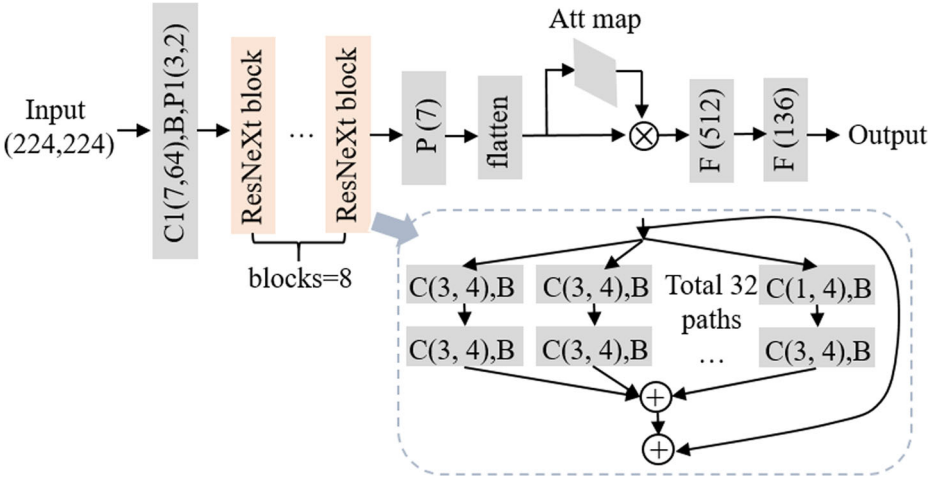

Figure 7 The structure of estimation sub-network. It consists of 17 convolution layers, including C_1, 8 stacked resnext blocks, a max pooling layer and a fully connected layer. In addition, we add an attention layer after the max pooling to enhance discriminating features. $C(k, n)$ indicates a convolutional layer, and its kernel size is k and output n feature maps, $P(k)$ represents a max pooling layer, and its kernel size is k, B represents batch-normalization. $\bigoplus$ indicates channel-wise sum operation and $\bigotimes$ is channel-wise Hadamard matrix product operation

exploit different CNN network compared with other papers. In the experiment section, we provide quantitative analysis.

4 Experiments

In this section, we conduct extensive experiments on several public datasets to evaluate our proposed model. Firstly, we describe implementation details in Section 4.1, such as training and testing datasets, error measurement metric, training procedure and so on. Section 4.2 shows ablation studies to evaluate the effectiveness of each component. Moreover, we also compare our method with state-of-the-art works in Section 4.3. Finally, we discuss issues with the current model.

4.1 Implementation details

Dataset In order to evaluate our model can tackle large unconditioned face images, we conduct experiments on the 300W competition dataset which covers large variation including different subjects, poses, illumination, occlusion, etc. It contains five databases: AFW, LFPW, HELEN, IBUG and 300W set [26, 27], where each image is annotated with 68 landmarks [25]. For training model, it is divided into training and testing parts following most established methods. The training set consists of images from AFW and IBUG dataset as well as the training subset of LFPW and HELEN. There are two testing sets, namely 300W public and 300W private testing sets. The 300W public set consists of images from the testing set of LFPW and HELEN. The 300W private set is the 300W set, which contains 300 outdoor images and 300 indoor images. To extend generalization of the model, we apply several times data augmentation on each training image, including randomly rotation, flipping, shifting and scaling.

Error measurements There are several common measurements for computing alignment error, for example, the mean distance between predicted landmarks and ground truth landmarks which normalized by the face's binocular distance, namely RMSE, and the failure rate of each landmark. The formula of RMSE is Eq. 7, where x_i^f, y_i^f is the ith predicted point, and $_i^g$, y_i^g is the ith ground truth point. The failure rate is the proportion of failed samples in all samples. In addition, we also plot Cumulative Error Distribution (CED) curves with respect to 68/51 facial landmarks on all datasets.

$$RMSE = \frac{\sum_{i=1}^{n} \sqrt{\left(x_i^f - x_i^g\right)^2 + \left(y_i^f - y_i^g\right)^2}}{d_{outer} N} \tag{7}$$

Experiment environment All experiments are implemented on the Ubuntu 16.04.3 system with Intel(R) Xeon(R) CPU E5-2630 v4 2.20GHz and two NVIDIA Corporation Device 1b02 GPUs and all deep convolution neural networks are designed based on the MXNET toolkit using the Python programming language. Our results are the average of multiple experiments.

Training procedure We train our model from scratching, network parameters are initialized by Xavier with the Gaussian function. The base learning rate of all networks is set to 0.0001 except networks applied STNs as it does not convergence. The initial learning rate of

learning-based method is between $1e-6$ and $1e-5$, and we set $3e-5$ to train related networks which achieves the best performance. We use Adam [13] stochastic optimization to update learning parameters and train each network 50 epochs. The loss function is the mean distance between predicted landmarks and ground truth landmarks. Here, it is necessary to point out that the setting of super-parameters is critical especially the base learning rate, as they can make the learned network outputs unreasonable results. We set these parameters based on experience and multiple times of debugging.

In the hand-crafted transformation, we first regress eight source points in the original image, then computing six-dimension affine transformation parameters θ. To convenience, we only align three points instead of eight points in the original image to corresponding positions in the mean face, as three point is enough. After experimental, we finally choose inter eyes corner and bottom lip center as source points; since these points contain full face region and the transformed image preserve more appearance similarity with the original image and the distortion is more reasonable.

Baselines We set two baseline models to compare results. The first and second baseline named HyN68 and ResNX68 respectively which directly outputs 68 facial landmarks without any warping operation based on the hyperface and ResNext network.

4.2 The effectiveness of spatial alignment network

Evaluating spatial transformation We propose spatial transformation module to get six-dimension affine transformation parameters as depicted in Section 3, so we first compare models with and without this module on 300W private/public set. As we propose two methods to get these parameters, then we compare similarities, differences and performance of these two methods in the implementation process. Finally, we compare the impact of network depth on performance. So, there are six models to compare, and results are shown in the first region of Tables 1 and 2. On the most challenging 300W private set, for adding the handcrafted transformation Hy8, HyN8+HyN68 model reduces the mean error by 0.7% (RMSE68) and 0.9% (RMSE51) compared with the first baseline HyN68, and Hy8+ReNX68 reduces the failure rate by 0.6% compared with the first baseline ReNX68. For extending the learning-based transformation Hy6, HyN6+HyN68 model reduces the mean error by 0.8% (RMSE68) and 1% (RMSE51) compared with the first baseline HyN68, but the reduced error of Hy6+ReNX68 model is not obvious, this is because that the estimation sub-network is far form the transformation sub-network and the learning process does not have its own loss function. In addition, on the 300W public set, the two proposed modules also help the original models achieve improved performance.

Compared these two modules, the handcrafted transformation is more stable than the learning based method, as it helps achieve stable improvement. But it is not as efficient as its opponent. To improve the learning-based transformation, we can add loss function to supervise the learning of transformation parameters, or add multi-transformation.

Further, we analyze the impact of network depth on performance. After getting transformed images, we make effort to regress 68 face landmarks, which is more difficult and needs more discriminating feature than regressing eight points. So, we apply ResNeXt instead of Hyperface to test its performance, and we provide experimental results for another 3 models, such as ReNX68, HyN8+ReNX68 and HyN6+ReNX68, as shown in Tables 1 and 2. We can see that deeper networks have lower error than shallow networks which validates our hypothesis.

Table 1 The result of facial landmark localization on 300W private test set based on our proposed baselines, where HyN and ReNX indicates hyperface and ResNeXt network respectively, the number after HyN/ReNX indicates the number of output values, the suffix Att means that an attention layer is added to the model

Model	RMSE68(%)	RMSE51(%)	Failure rate(%)
HyN68	7.2	6.4	10.0
HyN8+HyN68	6.5	5.5	6.6
HyN6+HyN68	6.4	5.4	6.9
ReNX68	5.1	4.2	2.6
HyN8+ReNX68	5.1	4	2.0
HyN6+ReNX68	5.6	4.5	4.5
HyN68_Att	6.2	5.4	4
ReNX68_Att	5	4.1	1.8
HyN8+ReNX68_Att	5	4	1.7
HyN6+ReNX68_Att	5.1	4.2	2.5

In addition, Hy8 and Hy6 stands for manual and learning transformation, respectively. RMSE is the mean distance between predicting landmarks and target landmarks and normalized by the binocular distance. The number after RMSE indicates the number of facial landmarks, including 68 facial landmarks and 51 facial landmarks, and 51 landmarks don't contain outer surface points of human face. The failure rate is the proportion of failed samples

Evaluating attention module We add an attention layer to our estimation sub-network to help understanding the image and select discriminating feature, as depicted in Section 3. We provide the result based on two baseline models without spatial transformation to see the importance of the attention mechanism directly. The results are shown in the second region of the Tables 1 and 2. In the shallow hyperface network, the HyN68_Att model reduces the

Table 2 The result of FLL on 300W public test set based on our model, where HyN and ReNX indicates hyperface and ResNeXt network respectively, the number after HyN/ReNX indicates the number of output values, the suffix Att means that an attention layer is added to the model

Model	RMSE68(%)	RMSE51(%)	Failure rate(%)
HyN68	7.2	6.4	10.0
HyN8+HyN68	5	4.2	2
HyN6+HyN68	4.9	4.2	1
ReNX68	3.7	2.9	0
HyN8+ReNX68	3.9	3	0.1
HyN6+ReNX68	4.1	3.2	0.2
HyN68_Att	4.7	4.1	0.7
ReNX68_Att	3.7	2.9	0
HyN8+ReNX68_Att	3.7	2.9	0
HyN6+ReNX68_Att	3.5	2.8	0.1

In addition, Hy8 and Hy6 stands for manual and learning transformation, respectively. RMSE is the mean distance between predicting landmarks and target landmarks and normalized by the binocular distance. The number after RMSE indicates the number of facial landmarks, including 68 facial landmarks and 51 facial landmarks, and 51 landmarks don't contain outer surface points of human face. The failure rate is the proportion of failed samples

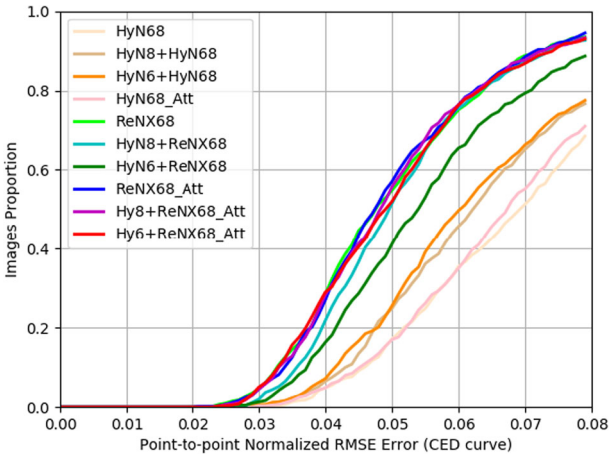

Figure 8 Fitting results on 300W private test set. The plots show the Cumulative Error Distribution (CED) curves with respect to the landmarks (68 landmarks)

failure rate by 6% after adding attention layer, and the deep ReNX68_Att model reduces the failure rate by 0.8%, so the attention layer is useful. But we find that the improvement in ResNeXt is not as great as in the hyperface, this is because the deep ResNeXt has extracted discriminating feature based on its multi-stacked convolution layers.

After evaluating the effectiveness of the spatial transformation and the attention module, we merged them to get the final model to further improve performance, and results are shown in the third region of the Tables 1 and 2.

In addition, we plot point-to-point Normalized RMS Error, as shown in Figures 8 and 9 which illustrating that the improvement is obvious after adding our proposed modules. We also show some original and transformed images in the Figure 10. We can see that the transformed images have near frontal viewpoint and have less background information.

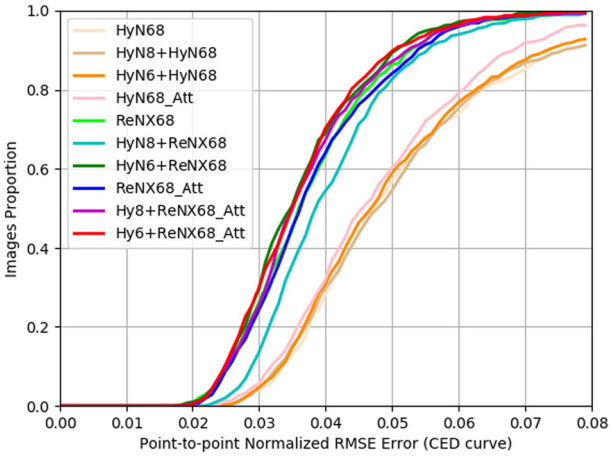

Figure 9 Fitting results on 300W public test set. The plots show the Cumulative Error Distribution (CED) curves with respect to the landmarks (68 landmarks)

Figure 10 The original (above) and transformed images (below)

4.3 The comparison with state-of-the-art

In this subsection, we compare our model with state-of-art methods on the challenging 300W private and public test set, and results are shown in Tables 3 and 4. Compared to the MDM [30] work which adopts deep CNN and RNN to locate facial landmarks, our model reduces the error by approximately 0.1 and 0.7. Compared to the DAN [14] which consists of three stages and similarity matrix. In the table, DAN(T1) and DAN(T3) represents result in the first and third stage. We can see that, our model performs better than DAN(T1) and is almost similar to DAN(T3). This means that if we extend our model into three stages, its performance can exceed the current best model. In addition, it is necessary to consider that our model only trained with 50 epoches without the fine-tune process that prove our model is robust and weakly dependent on hyper-parameters.

4.4 Discussion

Through extensive experiments, we find that considering spatial and appearance variation in unconstrained images is important, and our proposed model have achieved very good performance, as shown in Figure 11 which illustrates that the estimation is very close to the ground truth. But we also find some hidden issues. Firstly, our proposed two modules improves the performance of shallower CNN (Hyperface) a lot, while the improvement in a deeper and complex CNN (ResNeXt) is less obvious. So, in the future work, we will

Table 3 The result of FLL on 300W private set based on state-of-art methods

Models	RMSE68(%)
MDM [30]	5.1
DAN(T1) [14]	5.02
DAN(T3) [14]	4.3
Our	5

RMSE is the mean distance between predicting landmarks and target landmarks and normalized by the binocular distance

Table 4 The result of FLL on 300W public set based on state-of-art methods

Models	RMSE68(%)
MDM [30]	4.05
DAN(T1) [14]	–
DAN(T3) [14]	3.59
Our	3.5

RMSE is the mean distance between predicting landmarks and target landmarks and normalized by the binocular distance

Figure 11 Fitting results on 300W private test set (68 landmarks), red dot indicates the predicting position and blue dot indicates the ground truth. It is obvious that the predicting position is similar to the ground truth

Table 5 Percentage of images with fitting error of 68 landmarks less than the specified value based on our model

Model	< 0.02	< 0.03	< 0.04	< 0.05	< 0.06
HyNet8+Aff+ResNeXt68	0	0.32	0.27	0.51	0.75

Figure 12 Failed images on the 300W private test set (68 landmarks), red dot indicates the predicting position and blue dot indicates the ground truth. It is obvious that image under fairly extreme expression, pose, lighting and occlusion is still hard to align

consider how to upgrade them, for example, adding supervision in learning transformation parameters and designing a multi-scale attention module and so on.

Moreover, we further show image percentage with RMSE error less than 0.02, 0.03, 0.04, 0.05, 0.06 on the Table 5, and display the failed samples in the Figure 12 on 300W private test set. We can see that the sample with small error(0.02) is almost none, and the failed sample often occurs under extreme conditions. There may be three reasons for these problems. Firstly, our CNN model implements L2 loss which mainly focuses on large error images. Secondly, under very extreme conditions, the spatial and appearance variation on face images are still hard to analyze although implementing transformation as it is hard to predict their transformation parameters. Lastly, some annotations may be ambiguous and influenced by human factors, for example, people and software can't annotate low-resolution face images accurately. So, there's still a lot of effort to do, like designing a loss function mainly solving small errors and learning more flexible models to analyze images under very extreme conditions.

5 Conclusion

In this paper, we propose a Spatial Alignment Network to locating 68 landmarks on face images under unconstrained scenario. We propose a novel framework considering spatial and appearance variation, which consist of two modules. The first is the transformation sub-network converting spatial varied images to the canonical face and shape; we propose two methods to implement it, such as the hand-crafted and the learning-based method; the former method is stable, learning easily and straightforward but with low efficiency, while the latter is learnable and efficient but not that steady. The second module is the estimation sub-network to output 68 facial landmarks. In this module, we add an attention layer in the deep CNN to get more discriminating feature. Through extensive experiments, we validate the effectiveness of our proposed modules on several unconditioned datasets.

In the future, we will introduce other transformation methods and exploit other spatial and appearance information, like 3D and multi-scales/views information.

Acknowledgements This work is supported by National Science Foundation of China Grant #61672088 and #61790575, Fundamental Research Funds for the Central Universities #2018JBZ002. The corresponding author is Yidong Li.

References

1. Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Robust discriminative response map fitting with constrained local models. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3444–3451 (2013)
2. Bartz, C., Yang, H., Meinel, C.: Stn-ocr: a single neural network for text detection and text recognition (2017)
3. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 2930–2940 (2013)
4. Cai, Q., Gallup, D., Zhang, C., Zhang, Z.: 3d deformable face tracking with a commodity depth camera. In: Computer Vision - ECCV 2010, European Conference on Computer Vision, pp. 229–242. Proceedings, Heraklion (2010)
5. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2887–2894 (2012)

6. Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to scale: Scale-aware semantic image segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3640–3649 (2016)
7. Chu, Q., Ouyang, W., Li, H., Wang, X., Liu, B., Yu, N.: Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism (2017)
8. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation (2017)
9. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: European Conference on Computer Vision, pp. 484–498 (1998)
10. Dollar, P., Welinder, P., Perona, P.: Cascaded pose regression. IEEE **238**(6), 1078–1085 (2010)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition, pp. 770–778 (2016)
12. Jourabloo, A., Liu, X.: Pose-invariant 3d face alignment. In: IEEE International Conference on Computer Vision, pp. 3694–3702 (2016)
13. Kingma, D.P., Adam, J.B.a.: A method for stochastic optimization. Computer Science (2014)
14. Kowalski, M., Naruniec, J., Trzcinski, T.: Deep alignment network: A convolutional neural network for robust face alignment, pp. 2034–2043 (2017)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: International Conference on Neural Information Processing Systems, pp 1097–1105 (2012)
16. Li, H., Li, Y., Liu, W., Dong, H.: Coarse-to-fine facial landmarks localization based on convolutional feature. In: 2017 International Conference on Behavioral, Economic, Socio-cultural Computing (BESC), pp. 1–6 (2017)
17. Li, Y., Chang, M.-C., Farid, H., Lyu, S.: In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. arXiv:1806.02877 (2018)
18. Lin, C.H., Lucey, S.: Inverse compositional spatial transformer networks, pp. 2252–2260 (2016)
19. Liu, Y., Jourabloo, A., Liu, X.: Learning deep models for face antispoofng: binary or auxiliary supervision (2018)
20. Lv, J., Shao, X., Xing, J., Cheng, C., Zhou, X.: A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3691–3700 (2017)
21. Mo, K.: Spatial transformer network
22. Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2879–2886 (2012)
23. Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Trans. Pattern Anal. Mach. Intell. **PP**(99), 1–1 (2016)
24. Rashid, M., Gu, X., Yong, J.L.: Interspecies knowledge transfer for facial keypoint detection (2017)
25. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: A semi-automatic methodology for facial landmark annotation. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 896–903 (2013)
26. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge The first facial landmark localization challenge. In: IEEE International Conference on Computer Vision Workshops, pp. 397–403 (2014)
27. Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: database and results. Image Vis. Comput. **47**, 3–18 (2016)
28. Saragih, J.M., Lucey, S., Cohn, J.F.: Deformable Model Fitting by Regularized Landmark Mean-Shift. Kluwer Academic Publishers, Netherlands (2010)
29. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3476–3483 (2013)
30. Trigeorgis, G., Snape, P., Nicolaou, M.A., Antonakos, E., Zafeiriou, S.: Mnemonic descent method: a recurrent process applied for end-to-end face alignment. In: Computer Vision and Pattern Recognition (2016)
31. Tuzel, O., Marks, T.K., Tambe, S.: Robust face alignment using a mixture of invariant experts. In: European Conference on Computer Vision, pp. 825–841 (2016)
32. Xie, S., Girshick, R., Dollar, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks (2016)
33. Xiong, X., Torre, F.D.L.: Supervised descent method and its applications to face alignment. In: Computer Vision and Pattern Recognition, pp. 532–539 (2013)
34. Zhang, J., Shan, S., Kan, M., Chen, X.: Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In: European Conference on Computer Vision, pp. 1–16 (2014)
35. Zhang, Z., Luo, P., Chen, C.L., Tang, X.: Facial landmark detection by deep multi-task learning. In: European Conference on Computer Vision, pp. 94–108 (2014)
36. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)