Pattern Recognition 125 (2022) 108510

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Attentive occlusion-adaptive deep network for facial landmark detection

Muhammad Sadiq, Daming Shi*

College of Computer Science and Software Engineering, Shenzhen University, China

ARTICLE INFO

Article history:
Received 16 July 2020
Revised 20 December 2021
Accepted 22 December 2021
Available online 27 December 2021

Keywords:
Facial landmarks detection
Channel-wise attention
Spatial attention
Deep learning
Face alignment

ABSTRACT

To be very specific in this paper, an Attentive Occlusion-adaptive Deep Network, hereafter referred as AODN, is proposed for facial landmark detection, consisting of the geometry-aware module, attention module, and low-rank learning module. Facial Landmark Detection (FLD) is a fundamental pre-processing step of facial related tasks. Occlusion, extreme pose, different expressions and illumination are the main challenges in facial landmark detection related tasks. Convolutional Neural Network (CNN) based FLD methods have attained significant improvement regarding accurate FLD but, to deal with occlusion is still very challenging even for CNN. It is because; probably occlusion misleads CNN on feature representation learning. If faces are partially occluded, the localization accuracy will drop significantly. The role of attention in the human visual system is vital, and researchers proved its significance for the computer vision problem. Taking advantage of geometric relationships among different facial components and attention, we extended our already established Occlusion-adaptive Deep Network (ODN). We introduced the attention module consisting of Channel-wise Attention (CA) and Spatial Attention (SA) to improve its ability to deal with the occlusion and enhance feature representation ability simultaneously. The occlusion probability assists as adaptive weights of high-level features and minimizes the effect of the occlusion and assist in modelling the occlusion. Ablation studies prove the synergistic effect of each module. The summary of our trifold contribution is as follows: i) we introduced attention mechanism in our already established ODN model, to deal with occlusion more precisely, and get the rich feature representation to achieve better performance. ii) As per our best of knowledge, we are the pioneers to introduce CA and SA for FLD to model occlusion. iii) Our proposed methodology reduces the number of entire network parameters, which effectually decreases training time and cost. So, the proposed model is more suitable for scalable data processing. Experimental results prove the better performance of proposed AODN on challenging benchmark datasets.

1. Introduction

Facial points can be defined as the predetermined indication points on a given face graph. Mainly, these points are placed around the familiar components of the face, e.g., ear, eyes, nose, mouth, and chin. Usually, these points are located around or centre at some common facial components. The tasks related to facial analysis can differ based on numbers, types, the required quantity of landmarks, and the use of these landmarks. Localizing of these landmarks have been done for facial analysis tasks, it is going to take more attention of researchers during the last decade due to its importance. FLD is a key step for many facial analysis tasks, e.g., Facial action unit detection, face recognition, face expression detection, face frontalisation, head pose estimation as well as 3D face modelling [1] etc. The prospective behind FLD is to identify some precisely predefined key-markers on facial components. Still, there exist several challenges regarding this detection, e.g., Occlusion, illumination, expressions, extreme poses and so on.

Existing FLD methods can be categorised generally into 3 main groups: regression-based methods, template-based methods, and deep learning-based methods. The regression-based techniques learn the mapping from facial image appearance to landmark locations, different from template-based methods, regression-based methods do not usually build global shape models [2]. Regression-based methods predict all facial landmarks jointly, but the shape constraint and structured information is learned through the process [2].

Template-based methods leverage information about the facial appearance and global facial shape, controlling facial appearance and shape variations through statistical models [2]. Such models

* Corresponding author.
E-mail address: dshi@szu.edu.cn (D. Shi).

https://doi.org/10.1016/j.patcog.2021.108510
0031-3203/

Fig. 1. Examples of the occlusion by glasses, food, masks, hair, and hands, etc. From COFW dataset [14]. It can be easily observed that it is difficult to identify facial landmarks in presence of occlusion.

learn a parametric shape model through dataset which is already labelled, to model the changes in facial shape use Principal Component Analysis (PCA). Furthermore uses PCA to build global facial shape and facial appearance. All this process helps to refine the fitting algorithm. Some notable examples are: Active Shape Models (ASM) [3], 'Face detection, Pose estimation, and Landmark Localization' (FPLL) [4], Active Appearance Models (AAM) [5], and Discriminative Response Map Fitting (DRMF) [6]. The drawback of this kind of modes is, the reconstruction error effects whole face under occlusion, and as a result, all this leads model in hard circumstances, being unable to locate facial landmarks.

To solve computer vision problems, Deep Learning (DL) based methods are getting a prominent place. Inspired by the popularity and effectiveness of DL based methods, FLD researchers also started to apply DL techniques to solve FLD related problems. In comparison to conventional methods, DL based methods gained higher performance [7]. Lately, CNN based models gained a prominent place in DL based solutions to deal with FLD related tasks. But, to deal with occluded faces, is yet a challenge for CNN as well [8], it is because of the decrease in localizing accuracy due to occlusion. When the face is partially occluded, it is still challenging to improve localizing accuracy because occlusion probably deceives CNN during the learning of features.

To deal with the occlusion problem, the important is to identify the occlusion and model it. It is very challenging because its occurrence is random, irregular, and complex, as shown in Fig. 1. In literature, there exist several attempts to solve the occlusion problem. Wu and Ji [9] proposed a supervised regression method to update the probabilities steadily of landmark visibility at each iteration. In 2018 Xing et al. [10] proposed an occlusion dictionary into the already existing face appearance dictionary. The occlusion dictionary is learned and updated automatically in a data-driven manner. Liu et al. [11] proposed adaptive cascade regression besides of adaptive exemplar-based shape model to estimate the occlusion level of each landmark. Recently, in 2019 Zhu et al. [7] proposed BCNN-JBR. Recently Deng et al. [12] proposed integrated face bounding box prediction, 2D facial landmark localisation and 3D vertices regression under a unified multi-level face localisation task with a common goal: point regression on the image plane. In BCNN-JBR, each part of the face is treated with a separate pipeline. The objective is to share minimum information with other components pipelines to avoid correlative impact. As different facial components have a different number of facial points as well as different levels of hardness to predict facial points. It is challenging to calculate the unbiased hardness due to a lack of balance benchmark dataset. Like if during the prediction of mouth hardness if the dataset has more or less number of images having different expressions, occlusions, poses than other parts.

Earlier, we proposed Occlusion-adaptive Deep Network (ODN) [13] for the FLD task to deal with occlusion. Originally ODN is further divided into three modules, i.e., Distillation Module (DM), Geometry-aware Module (GM), and Low-rank learning Module (LM). We used DM to model occlusion based on the probabilities of the high-level features. The objective of GM is to identify the relations between features of different facial components. Furthermore, the DM's aim is to recover the missing features.

ODN outer performs than previous methods, but the current results are not exactly as per our expectation. After investigating the reasons in detail, we identified that ODN doesn't have enough rich feature representation and not suitable to model the occlusion more precisely. It is because Real-world images are affected by appearance and spatial distortion. It is because of the irregular position of the camera according to the scene. It alters the dimensions, which causes the geometry of the scene, and gradually performance declines [15].

To address the problem of spatial distortion and obtain rich feature representation, we introduced the attention mechanism into our already established ODN model. To be very specific, we replaced the distillation module with proposed attention module, consist of Channel-wise Attention (CA) and Spatial Attention (SA). CA focus on "what" is relevant and meaningful in a given facial image, while the purpose of SA is to guide the network "where" to focus.

The summary of our trifold contribution is as follows: i) we introduced attention mechanism in our already established ODN model, to deal with occlusion more precisely, and get the rich feature representation to achieve better performance. ii) As per our best of knowledge, we are the pioneers to introduce CA and SA for FLD to model occlusion. Experimental results prove the better performance of proposed AODN on challenging benchmark datasets.

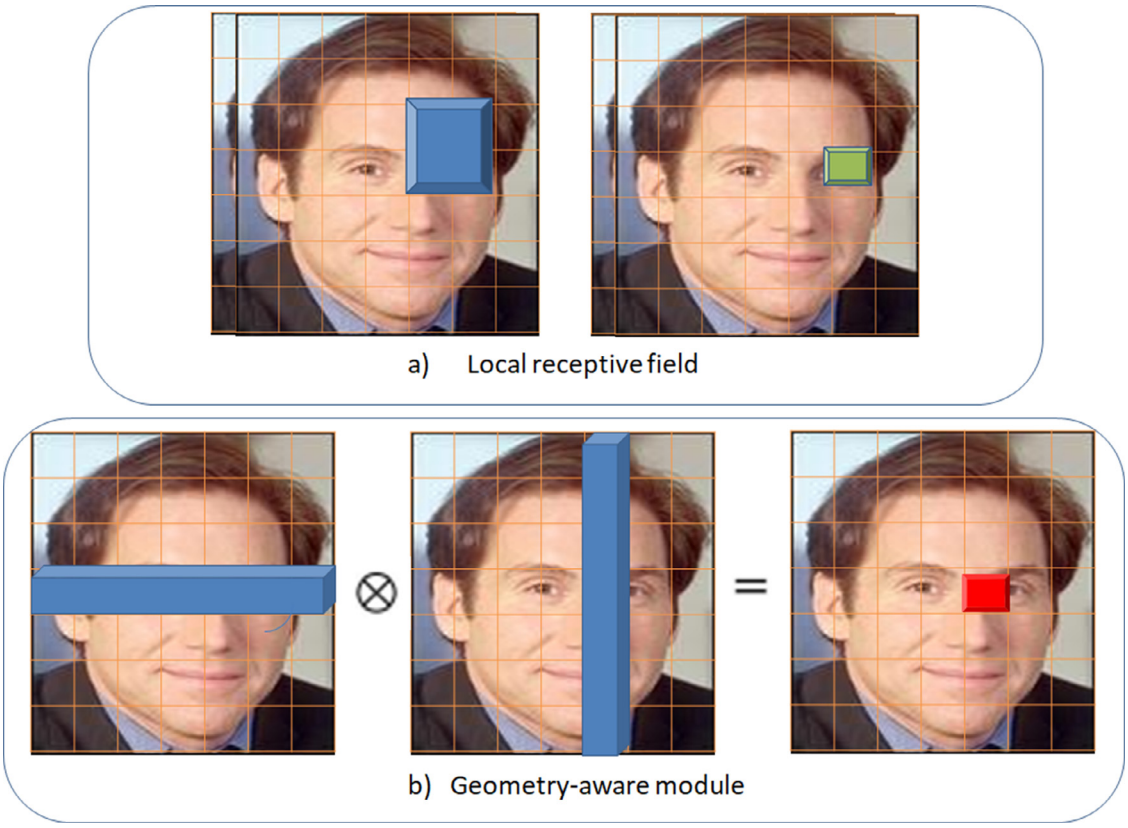

Fig. 2. Comparison of structure-aware module and local receptive field on capturing geometric relations of facial images.

iii) Our proposed methodology reduces the number of entire network parameters, which effectually decreases training time and cost. So, the proposed model is more suitable for scalable data processing.

The remainder of the paper is organized as follows. Some basic preliminary and background information is given in Section 2. We elaborate our proposed Attentioned Occlusion-adaptive Deep Network (AODN) solution in Section 3. Mathematical optimization is presented in Section 4. Detailed experiments of our proposed framework are spelled-out in Section 5. Section 6 draws the conclusion of this paper.

2. Background and preliminary

The primary objective of FLD is to identify some predefined key-points of a given facial image on facial components. Although FLD gained significant success during the last decade, the occlusion is still a significant barrier to achieve it perfectly. To solve occlusion, it is essential to determine how to model occlusion. From the facial appearance, model the occlusion is very difficult because of its irregular occurrence, randomness, and complexity.

In human perception, the role of attention is vibrant. The human visual system usually extracts glimpses in parts and focuses on a specific part selectively to capture a fine visible structure. If we carefully observe the human visual system, due to a vast quantity of connections of cones, only foveolar visual acuity approaches up to one hundred percent [16]. In CNN, either it is space or time, the convolutional process just processes the local neighbor dependencies. The long-range dependencies just only can be captured by having a repetition of the same process. Repetition of the same process affects accuracy, time and cost. However, for FLD related tasks, the geometric relation between different facial components belongs to long-range dependencies. Attention has a significant role in capturing long-range dependencies [17]. Usually, the prospective behind attention is to tell network precisely where to focus, by calculating the response for a particular location as a weighted sum of the features at all positions. In this section, we will discuss the structure of ODN and its weaknesses, followed by some discussion about attention mechanisms.

2.1. Already established occlusion-adaptive deep network (ODN)

Overall ODN structure can be divided into three modules, as shown in Fig. 3: The LM followed by the DM and GM A shared structure matrix is generated by GM and DM to help LM to recover missing features. The purpose of the GM is to capture the geometric structure of the facial image, as mentioned in Fig. 2. The use of DM is to model occlusion probability. In ODN, occlusion probability is based on high-level features. In regular scenarios, irregular positions of the camera affect facial images by spatial and appearance variations because the images are collected in the wild [15].

To be very specific, to obtain ODN, we modified the ResNet-18 [18] structure. We edited its last residual block to get the GM, DM, and LM. In ODN, the obtained feature map from previous residual block fed into the DM and GM. The objective is to get clean feature representation and geometric information, respectively. GM and DM have two sperate pathways subnetworks. It is the same as quadratic kernel expansion to obtain rich feature representation. As already mentioned, the distillation module also consists of two different sub-pathways networks.

We observe that the distillation module doesn't have enough rich feature representation and is not able to focus on occluded parts. Motivated by the success achieved by deep learning, we propose in this paper a novel framework AODN to address this issue. In AODN, we introduce the attention module instead of the distillation module. Experimental results show that our model per-

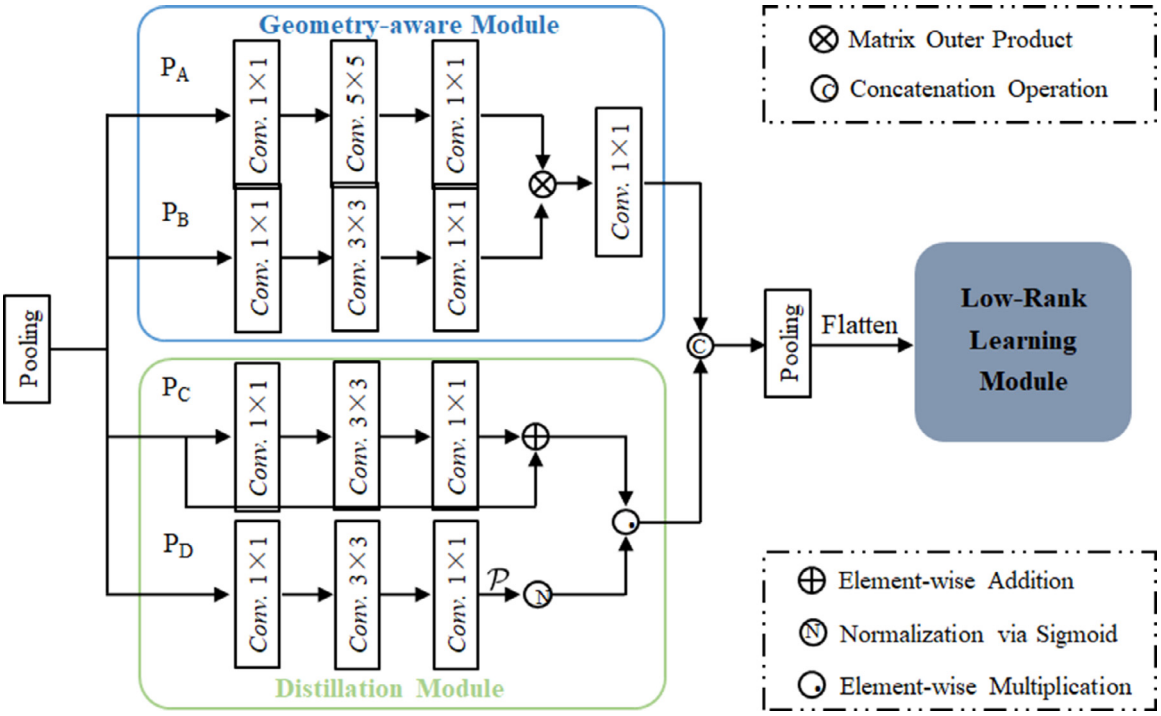

Fig. 3. Structure of Occlusion-adaptive Deep Network [13].

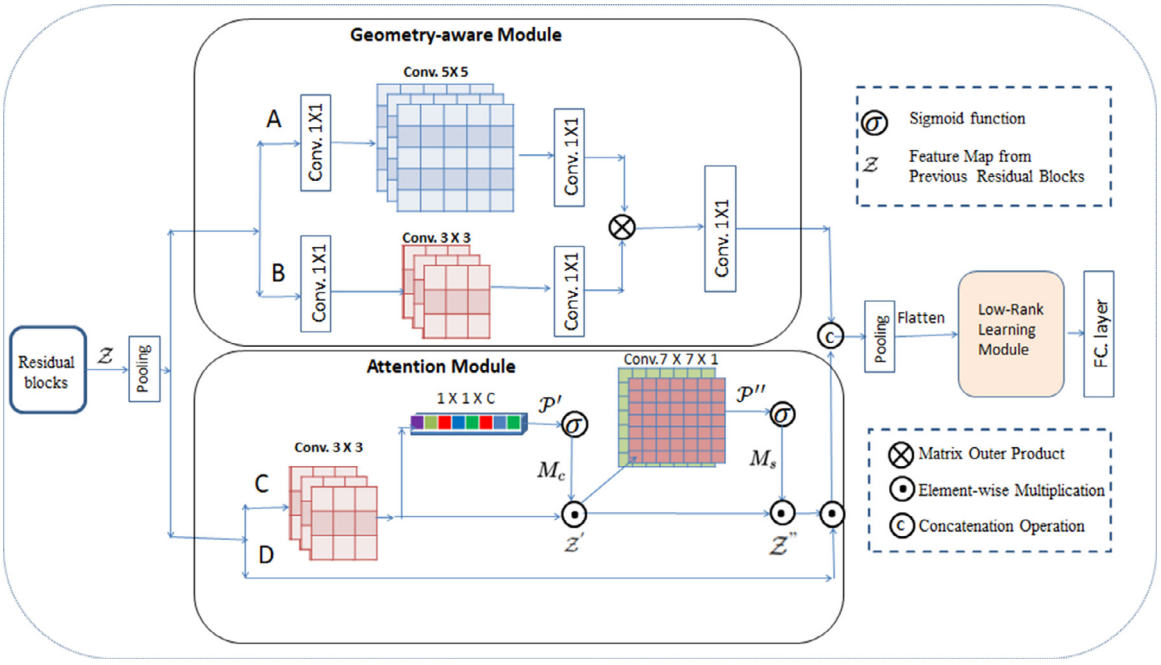

Fig. 4. The diagram of the proposed Attentive Occlusion-adaptive Deep Network.

forms better than current state-of-the-art methods on benchmark datasets.

2.2. Attention for facial landmarks detection

Attention has a very significant role to capture long-range dependencies [17]. Usually, the prospective behind attention is, to tell network precisely where to focus, by calculating response for a particular location as a weighted sum of features at all positions [19,20].

Attention has a vital role in the human visual system. To improve the performance of CNN by using attention, in literature there exist several attempts. Li et al. [15] proposed the Spatial Alignment Network (SAN) based on the hand-crafted method and learning-based method. The basic target problem of SAN is to deal, appearance and spatial variations. But the problem with SAN is, if it uses a handcrafted method, the efficiency is very low. If it uses a learning base method, it is not steady. To improve network performance AAN [21] and JAA-Net [22] also use spatial attention for FLD. As we already discussed, the purpose of SA is to tell network 'where' to focus, but still, the network is not aware of 'what'

to focus. In image classification tasks CA and SA [16,17,23,24] together achieved significant improvement. In ADN [24], a separate network is used to model occlusion. Attention is used to get features as an additional layer. Different from ADN, we used CA and SA to model occlusion and get rich feature representation simultaneously. Furthermore, we used element-wise multiplication instead of simple addition to identify occluded parts and background probability maps. The motivation behind element-wise multiplication is to get more vigilant features and distinguish more precisely between regular components and occluded parts. We incorporate CA and SA in our already established ODN [13] to improve the performance (Fig. 4).

3. Attentive occlusion-adaptive deep network(AODN)

As already mentioned, deep learning-based methods have a significant role to deal with computer vision issues. Mostly deep learning base methods use CNN and have proved its significance in deep learning-based models for FLD. In CNN, either it is space or time, the convolutional operation just processes the local neighbour dependencies. The long-range dependencies just only can be captured by having a repetition of the same process. Repetition of the same process affects accuracy, time, and cost. However, for the task of FLD, the geometric relation between various facial components belongs to long-range dependencies. Usually, the prospective behind attention is to tell network precisely where to focus, by calculating the response of a particular location as a weighted sum of features at all positions [17].

The features of the occluded region are filtered by attention. The non-existence of some features doesn't mean that, the face doesn't have those features. This could be a biased explanation of the model. Most of the face features co-occur or are co-related with each other. It can help to recover missing features because some features have position relation, proximity or symmetry. So the existence of some features helps to recover missing features or leads towards other features.

As defined in Fig. 4, the overall structure of AODN can be divided into three different modules and four sub-networks. The geometry-aware module is based on two sub-networks A and B. The attention module consists of sub-network C and D. The sub-network C exploits a residual block to implement CA and SA, respectively. The main objective is to model occlusion more precisely and get the rich representation of facial features. The aim of sub-network D is to avoid input signal decay, to obtain stable features. Furthermore, the output of C and D is integrated by element-wise multiplication, to assign small weights to the background and occluded parts. Finally, as output from sub-network C and sub-network D, for the holistic face, a weighted feature map (clean features) can be obtained. To make it more sparse during optimization, we take advantage of the L_1 regularisation technique on $\mathcal{P}^{'}$, and $\mathcal{P}^{''}$ to make it more robust, and sparse during optimization. Finally, the output from both modules (geometry-aware and attention) is concatenated to obtain a high dimensional single feature map, to obtain hybrid feature map of the facial graph. Furthermore, these hybrid feature maps are used as input of the low-rank learning module after down-sampling and flattering. Given the training set $\left\{\left(I_i, \check{S}_i\right)\right\}$ can be learned by (1).

$$\min \frac{1}{N}\sum_{i=1}^{N}\left\|\check{S}_i - S\right\|_F^2 + \beta Rank(\mathcal{M}) \tag{1}$$

Where $\check{S}$ and S represents ground-truth and corresponding prediction. $\check{S} = \{s_1, s_2, \ldots . s_L\}$ and $S = W_{fc}^T \mathcal{M}^T \mathcal{X}$. The outputs of the geometry-aware module is denoted as $\mathcal{X}$. L and s are the numbers of landmarks and facial landmarks, respectively. β is used as a regularization factor to adjust the rank of $\mathcal{M}$. W_{fc} means, the parameters of a fully connected layer. The AODN trained in an end-to-end manner the same as ODN.

3.1. Attention module

As already discussed, feature representation has a significant role to model occlusion more precisely and learn missing features more proficiently. The effectiveness of CA and SA for image classification related tasks has already been determined by Chen et al. [16], Woo et al. [17], Park et al. [23]. Inspired by their work, we incorporated CA and SA to obtain rich feature representation and model occlusion more accurately. The detailed structure and implementation details of CA and SA are shown in Figs. 5, and 6, respectively. In AODN attention module consists of sub-network C, and sub-network D. The sub-network C exploits a residual block to implement CA and SA, respectively. The main goal is to model occlusion more accurately and get the rich representation of facial features. The aim of sub-network D is to avoid input signal decay, to obtain stable features. Overall refined feature map can be defined as:

$$\mathcal{F} = \mathcal{Z}^{''} \bullet \mathcal{Z} \tag{2}$$

where $\mathcal{F}$ represents the final feature map after merging the residual map and the attention map. The $\mathcal{Z}^{''}$ is a feature map obtained through the attention process, and $\mathcal{Z}$ represents the feature map of previous residual blocks.

3.2. Channel-wise attention and spatial attention

Usually, researchers increase the width or depth of the corresponding network to get better feature representation during network engineering. In deep networks, the increase in network parameters affects the efficiency in terms of cost and time. Furthermore, the assembling of features accurately in reverse order is also a questing mark. To deal with this issue [16,23] used attention and proved its effectiveness for image classification tasks. Same as the human visual system in network engineering, attention helps to get rich feature representation. Furthermore, attention tells the network about the specific area that needs to be focused on. To simplify, we also used CA to guide the network 'what' is meaningful in a given facial image, and SA guide the network 'where' to focus. The objective behind this attempt is to ensure the sensitivity of network to informative features. As already discussed, channel attention can be obtained by exploiting inter-channel relation to get similar features for the specific landmark. To simplify, CA and SA can be written, as mentioned in (3) and (4), respectively.

$$\mathcal{Z}^{'} = M_c(\mathcal{Z}) \bullet \mathcal{Z} \tag{3}$$

M_c, $\mathcal{Z}$ denotes the channel-wise attention map, and the feature map of prior residual blocks, respectively.

$$\mathcal{Z}^{''} = M_s(\mathcal{Z}^{'}) \bullet \mathcal{Z}^{'} \tag{4}$$

$\mathcal{M}_s$ is the spatial attention map and $\mathcal{Z}^{'}$ is the feature map obtained by channel-wise attention.

The objective of CNN is to extract the features from a given image. If an image $W \times H \times 3$ passes over convolutional layer with C channels. CNN uses filters to scan the given image and produce a $\acute{W} \times \acute{H} \times C$ feature map as output, it can be input of further convolutional layers. Usually, filters act as a pattern indicator, i.e., high-level filters indicate the syntactic fashion like objects, parts, and low-level filters indicate the lower-level visual fashion like edges and corners, etc. If we closely observe the CNN, it abstracts the features by stacking of multiple layers through a pyramid of visual intellection. To simplify, we can say image features of CNN are multi-layer, channel-wise, and spatial. In the context of FLD, the selection of semantic features on-demand can be viewed as CA.

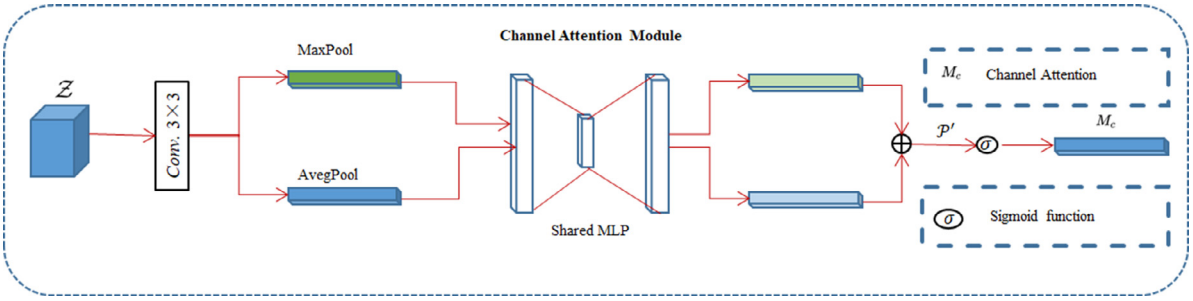

Fig. 5. A detailed implementation structure of channel attention.

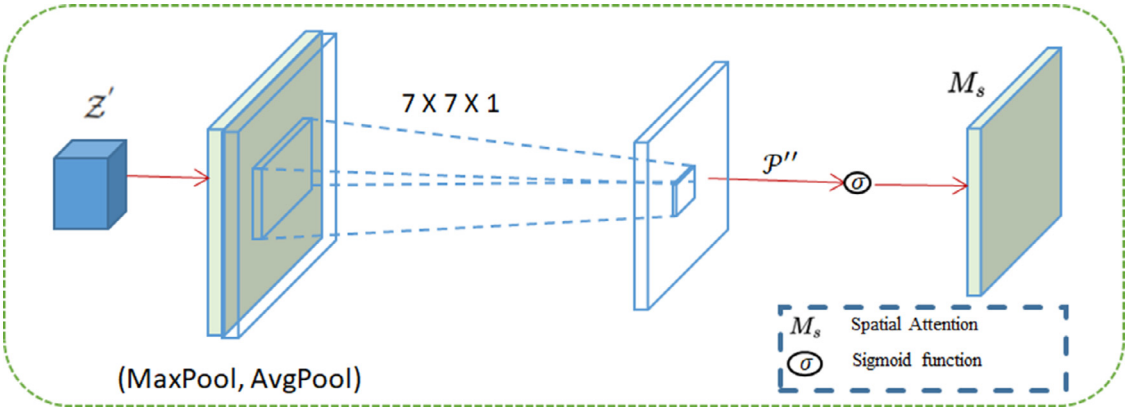

Fig. 6. A detailed implementation structure of spatial attention.

3.2.1. Channel-wise attention

In CNN, each channel of feature map performs as a feature detector. So, by exploiting the relationship between channels, the channel attention map has been formed. As already discussed, the CA aims to focus "what" is relevant and meaningful in a given facial image. The purpose of SA is to guide the network "where" to focus. The detailed structure of CA is as per Fig. 6. To get a channel attention map, we squeezed the spatial dimension of the input feature map. To get distinct features of the given object to refine CA, we used max-pooling along with average-pooling. As max-pooling guides network more precisely towards more distinct features. Mathematically, the channel attention map can be as per Eq. (5).

$$\begin{aligned} M_c(\mathcal{Z}) &= \sigma\left(MLP(AvgPool(\mathcal{Z})) + MLP(MaxPool(\mathcal{Z}))\right) \\ &= \sigma\left(W_1(W_0(\mathcal{Z}^c_{avg})) + W_1(W_0(\mathcal{Z}^c_{\max}))\right) \end{aligned} \tag{5}$$

We accumulated spatial information of feature maps through max-pooling and average-pooling to obtain channel attention. Later, two separate context descriptor: $\mathcal{Z}^c_{Max}$ and $\mathcal{Z}^c_{Avg}$, represents max-pooled and average-pooled features correspondingly. As mentioned in Fig. 5, a shared network with both descriptors generates channel attention map $M_c \in \mathcal{R}^{1\times1\times C}$. The shared network consists of Multi-Layer Perceptron (MLP) with one hidden layer. We used hidden activation size $\mathcal{R}^{1\times1\times\frac{C}{r}}$ to reduce parameter overhead, where r stands for reduction ratio. We combined the output feature vectors by element-wise addition after applying the shared network to each descriptor. Where σ means the sigmoid function and $W_0 \in \mathcal{R}^{\frac{C}{r}\times C}$ and $W_1 \in \mathcal{R}^{C\times\frac{C}{r}}$, where W_0, W_1 means the shared weights of MLP.

3.2.2. Spatial attention

We used the inter-spatial relationship of features to obtain spatial attention map. The objective of SA is to direct the network 'where' to emphasize more specifically, it is corresponding to CA. We computed SA by applying pooling beside the channel axis. The detailed implementation structure of SA is illustrated in Fig. 6. We used max-pooling and average-pooling besides channel axis and concatenated both to obtain feature descriptors. Furthermore, to attain spatial attention map $M_s(\mathcal{Z}) \in \mathcal{R}^{H\times W}$ we use a convolutional layer to direct the network 'where' to emphasize. We combined channel information by using max-pooling and average-pooling operations, to generate two 2D maps: $\mathcal{Z}^s_{avg} \in \mathcal{R}^{H\times W\times1}$ and $\mathcal{Z}^s_{\max} \in \mathcal{R}^{H\times W\times1}$.

$$\begin{aligned} M_s(\mathcal{Z}) &= \sigma\left(f^{7\times7}([AvgPool(\mathcal{Z}); MaxPool(\mathcal{Z})])\right) \\ M_s(\mathcal{Z}) &= \sigma\left(f^{7\times7}([\mathcal{Z}^s_{avg}; \mathcal{Z}^s_{\max}])\right) \end{aligned} \tag{6}$$

3.3. Fundamental relationship between three modules

Our proposed attentive occlusion-adaptive framework has a very close-knit relationship among three modules, i.e., geometry-aware module, attention module, and low-rank learning module. As mentioned earlier [13], the human brain's visual processing is involved with two streams: the ventral stream and the dorsal stream. The ventral stream takes charge of the discrimination and recognition of objects while the dorsal processes the object's spatial location information. Similar to this mechanism, our proposed AODN is related to two primary information: occlusion-awareness and geometric relationships. To be specific, there exist invariable solid geometric relationships among different facial components, e.g., symmetry, proximity, position relation and so on, which the geometry-aware module can capture.

On the other hand, occlusion regions and irrelevant information from the background can be filtered by the proposed attention module. Same as the human visual system in network engineering, attention helps to get rich feature representation. Furthermore, attention tells the network about the specific area that needs to be focused on. To simplify, we also used CA to guide the network 'what' is meaningful in a given facial image, and SA guide the net-

work 'where' to focus. The objective behind this attempt is to ensure the sensitivity of network to informative features. Some lost information from one component can be speculated via other components according to the geometric characteristics. In addition, the relation of opposite and complementary between attention module and low-rank learning module is benefited to the feature learning of face. Although the hybrid features can improve performance, the hybrid features are not exhaustive and complete feature representation of holistic face because the attention module filters the features of occluded regions. From the above, the structural relationship among the three modules boosts our proposed AODN to deal with the occlusion problem.

4. Optimization

The following minimization problem is to formulate AODN mathematically:

$$\min \frac{1}{N}\sum_{i=1}^{N}\left\|\breve{S}_i - S_i\right\|_F^2 + \beta\mathcal{R}ank(\mathcal{M}) + \gamma\|\mathcal{M}\|_F^2 + \alpha\|\mathcal{W}_c\|_F^2 + \lambda\left\|\mathcal{W}_{fc}\right\|_F^2 + \eta^{'}\left\|\mathcal{P}_i^{'}\right\|_F^1 + \eta^{''}\left\|\mathcal{P}_i^{''}\right\|_F^1, \tag{7}$$

Where $\mathcal{S} = \mathcal{F}_{AODN}(\mathcal{I}; \mathcal{W}_{fc}; \mathcal{M})$, and $\mathcal{F}_{AODN}(.)$ is our proposed AODN. $\mathcal{W}_c$ represents the parameter set of convolutional layer, and $\mathcal{W}_{fc}$ represents the fully connected layer. $\mathcal{M}$ is the parameter set of LM. Frobenius norms control the shrinkage of all three given parameter sets with the connected parameters $(\alpha, \gamma, \lambda)$, respectively. The single-channel Feature map $\mathcal{P}^{'}$, and $\mathcal{P}^{''}$ from attention module with parameter η is imposed by L_1.

Usually, the purpose of nuclear norm $\|\mathcal{M}\|_*$ is to resolve the problems related to low-rank learning. It provides the tightest lower bound among all convex lower bounds of the rank function. So, Eq. (7) can be re-written as:

$$\min \frac{1}{N}\sum_{i=1}^{N}\left\|\breve{S}_i - S_i\right\|_F^2 + \beta\|\mathcal{M}\|_* + \gamma\|\mathcal{M}\|_F^2 + \alpha\|\mathcal{W}_c\|_F^2 + \lambda\left\|\mathcal{W}_{fc}\right\|_F^2 + \eta^{'}\left\|\mathcal{P}_i^{'}\right\|_F^1 + \eta^{''}\left\|\mathcal{P}_i^{''}\right\|_F^1, \tag{8}$$

By applying the property of circularity of trance and the definition of nuclear norm as per [25] we can obtain

$$\begin{aligned}\|\mathcal{M}\|_* &= \mathrm{tr}(\sqrt{\mathcal{M}^{T\mathcal{M}}}) \\ &= \mathrm{tr}\left(\sqrt{\left(U\Sigma V^T\right)^T\left(U\Sigma V^T\right)}\right) \\ &= \mathrm{tr}(\sqrt{V\Sigma^2V^T}) \\ &= \mathrm{tr}(\sqrt{VV^T\Sigma^2}) \\ &= \mathrm{tr}(\sqrt{\Sigma^2}) \\ &= \mathrm{tr}(|\Sigma|)\end{aligned} \tag{9}$$

We apply Singular Value Decomposition (SVD) [26] to obtain U, Σ, and V. We can find subgradient as given below in (10), because $|\ \Sigma\ |$ is not differentiable on every point in its domain.

$$\begin{aligned}\frac{\partial\|\mathcal{M}\|_*}{\partial\mathcal{M}} &= \frac{\partial tr(|\ \Sigma\ |)}{\partial\mathcal{M}} = \frac{tr(\partial\ |\ \Sigma\ |)}{\partial\mathcal{M}} \\ &= \frac{tr(|\ \Sigma\ |\ \Sigma^{-1}\partial\Sigma)}{\partial\mathcal{M}}.\end{aligned} \tag{10}$$

We know $\mathcal{M} = U\Sigma V^T$ and $\partial\mathcal{M} = \partial U\Sigma V^T + U\partial\Sigma V^T + U\Sigma\partial V^T$ Hence, $U\Sigma\partial V^T = \partial\mathcal{M} - \partial U\Sigma V^T - U\Sigma\partial V^T$. By multiplying U^T on the left side and V on the right side of (11), we will obtain

$$\partial\Sigma = U^T\partial MV - U^T\partial U\Sigma - \Sigma\partial V^TV, \tag{11}$$

In (11) U represents the unitary matrix, i.e, $U^TU = I$. where I is an identity matrix. The rank of the second term in (11) can be computed as below:

$$\begin{aligned}tr(U^T\partial U\Sigma) &= tr((U^T\partial U\Sigma)^T) = tr(\Sigma^T\partial U^TU) \\ &= -tr(\Sigma U^T\partial U) = -tr(U^T\partial U\Sigma),\end{aligned} \tag{12}$$

Equation (12) indicates that $tr(\Sigma U^T\partial U), tr(\Sigma - \Sigma\partial V^TV) = 0$. So we can obtain $tr(\partial\Sigma) = tr(U^T\partial\mathcal{M}V)$. Hence, substituting it into (11) we can have

$$\begin{aligned}\frac{\partial\|\mathcal{M}\|_*}{\partial\mathcal{M}} &= \frac{tr(|\ \Sigma\ |\ \Sigma^{-1}\partial\Sigma)}{\partial\mathcal{M}} \\ &= \frac{tr(|\Sigma|\Sigma^{-1}U^T\partial MV)}{\partial\mathcal{M}} \\ &= \frac{tr(V\ |\ \Sigma\ |\ \Sigma^{-1}U^T\partial M)}{\partial\mathcal{M}} \\ &= (V\ |\ \Sigma\ |\ \Sigma^{-1}U^T)^T \\ &= U\Sigma^{-1}\ |\ \Sigma\ |\ V^T,\end{aligned} \tag{13}$$

in the given objective function, in this way, we can attain the gradient of the rank function.

$$\frac{\partial\|\mathcal{P}\|_1^F}{P_k} = \begin{cases}+1 & P_k > 0 \\ -1 & P_k < 0 \\ [+1, -1], & P_k = 0\end{cases} \tag{14}$$

where P_k is kth element in $\mathcal{P}$, and $\mathcal{P}$ be $\mathcal{P}^{'}$ or $\mathcal{P}^{''}$. Hence, we can say that it is a directed acyclic graph and the gradients of the regression loss (e.g., L_2 loss) can be optimized by back-propagation. Furthermore, all parameters can be, in an end-to-end fashion.

Table 1
The NRMSE $(\times 10^{-2})$ comparison results on common set and full set of 300W.

Method	Year	Common Set	Fullset
Seq_MT [33]	2018	4.20	4.90
PCD-CNN [32]	2018	3.67	4.44
ODN [13]	**2019**	**3.56**	**4.17**
ADN [24]	2019	3.52	4.14
LGSA [35]	2020	3.36	4.06
3FabRec [36]	2020	3.36	3.82
ADC [37]	2020	**2.83**	4.23
RetinaFace [12]	2020	**2.74**	3.91
AODN	**2020**	3.27	**3.76**
AODN+	**2020**	3.14	**3.63**

Table 2
The NRMSE $(\times 10^{-2})$ comparison results on Challenging set of 300W.

Method	Year	Challenging set
DSRN [41]	2018	9.68
SBR [42]	2018	7.58
SAN [34]	2018	7.55
ODN [13]	2019	6.67
ADN [24]	2019	6.60
ADC [37]	2020	7.04
RetinaFace [12]	2020	6.83
AODN	**2020**	**6.38**
AODN+	**2020**	**5.91**

5. Experiments

In this section, we broadly assess the performance of the proposed framework on different benchmark datasets under various settings for the task of FLD. First in Section 5.1, we present some initial details about the benchmark datasets and used experimental settings. Then, Section 5.2, the evaluation metric, and employment details for the training of AODN. Following that, we investigate the effect of several system parameters as well as the contribution of the different components to the FLD performance. In Section 5.3,

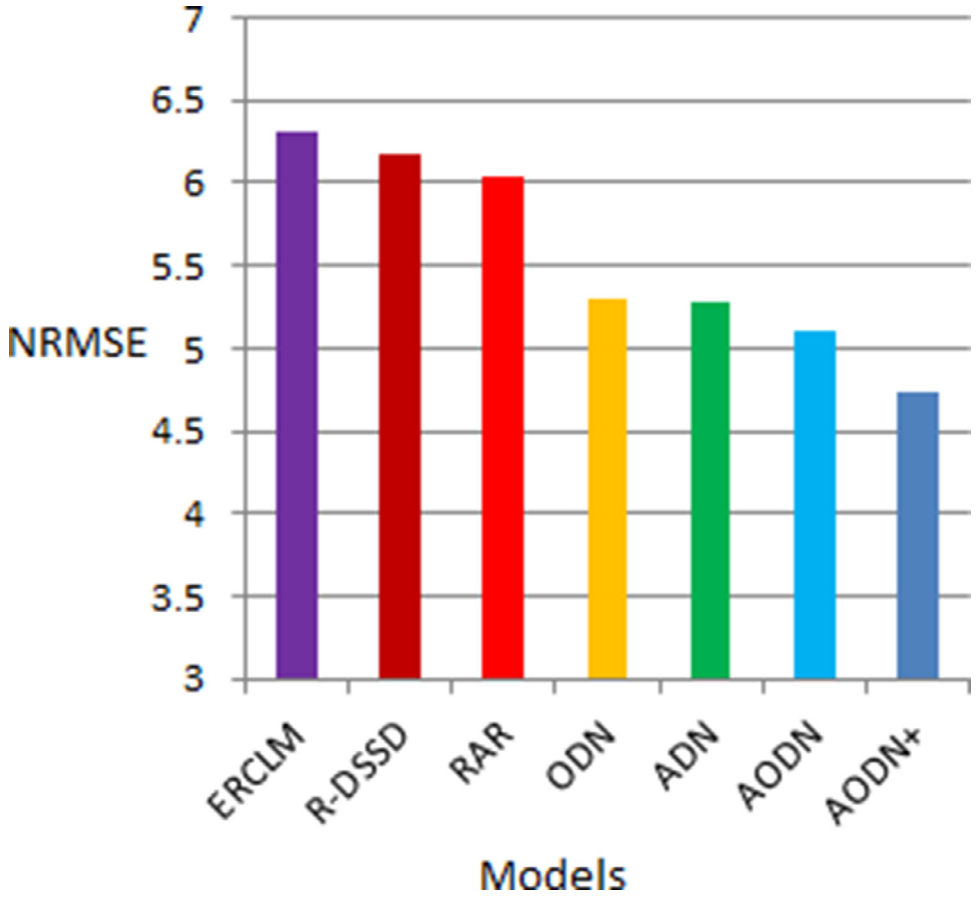

Fig. 7. The NRMSE $(\times 10^{-2})$ comparison results on COFW dataset.

Table 3
The NRMSE $(\times 10^{-2})$ comparison results on 300VW dataset for all three categories.

Method	Year	Cat. 1	Cat. 2	Cat. 3
TSTN [43]	2017	5.36	4.51	12.84
AAN [21]	2018	5.03	4.82	7.98
ADN [24]	2019	4.75	4.34	6.72
AODN	**2020**	**4.69**	**4.26**	**6.67**
AODN+	**2020**	**4.52**	**4.09**	**6.26**

Table 4
The NRMSE $(\times 10^{-2})$ comparison results on AFLW Dataset.

Method	Year	AFLW-Full	AFLW-Frontal
SAN [34]	2018	1.91	1.85
DSRN [41]	2018	1.86	-
ODN [13]	2019	1.63	1.38
3FabRec [36]	2020	1.84	1.59
DCSD [44]	2020	1.62	-
RetinaFace [12]	2020	1.41	**1.11**
AODN	**2020**	**1.38**	1.13
AODN+	**2020**	**1.12**	**1.03**

we present the ablation study to validate our framework. We resized and cropped all images (224×224) and perform a flip, scale, rotation tasks and translation to do the data augmentation for the training set. Same as ODN, we also pre-trained all our models on the ImageNet dataset [27].

Table 5
The ablation analysis on 300W Challenging set.

Model	NRMSE
BRNet	7.21
BRNet+GM+LM	6.90
BRNet+GM+AM	6.72
BRNet+AM+LM	6.68
BRNet+GM+LM+AM(r=8), (without L1)	6.51
BRNet+GM+LM+AM(r=32)	6.47
BRNet+GM+LM+AM(r=64)	6.46
BRNet+GM+LM+AM(r=16)+EWA	6.46
BRNet+GM+LM+AM(r=16)	6.44
BRNet+GM+LM+AM(r=8)+ EWA	6.41
BRNet+GM+LM+AM(r=8)	**6.38**

5.1. Datasets

To broadly assess the performance of the proposed framework on different benchmark datasets under various settings for the task of FLD, we use the ensuing diverse standard datasets: 300W [28], AFLW [29], COFW [14], Menpo [30], and 300VW [31]. All these datasets are benchmark datasets and publicly available for research purposes. We compare outcomes of our method with contemporary methods [7,13,21,24,32–34]. In the first step, we trained our model on a 300W training set and tested our model on other datasets. In 2nd phase we trained our model on the Menpo training set and tested on other datasets. To mention the result separately, we named *AODN+* to model, trained on Menpo.

- 300W is a re-known, widely used, and publically available dataset for FLD, to measure the effectiveness of the method.

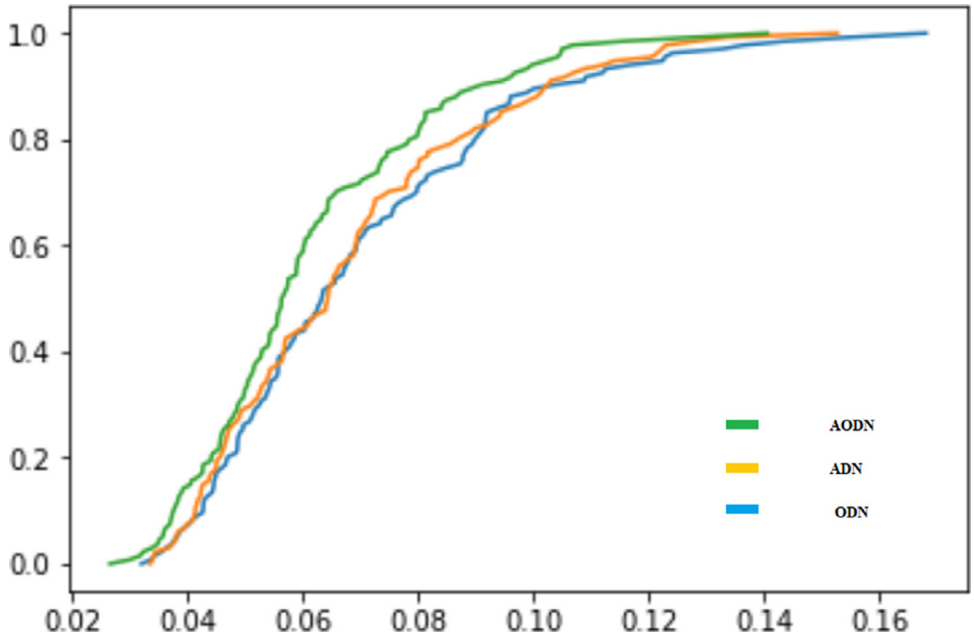

Fig. 8. The CED curve on 300W challenging set.

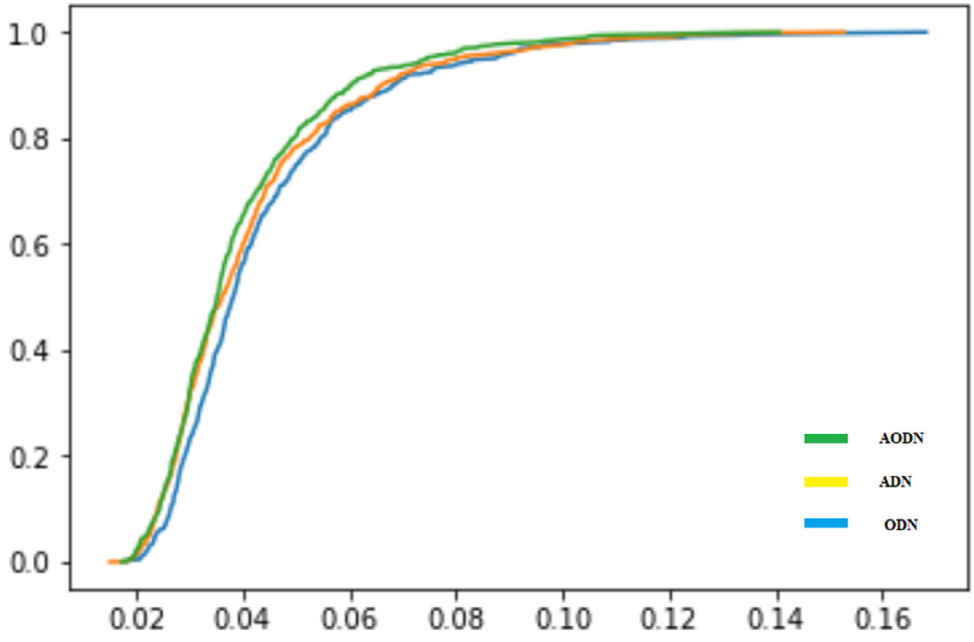

Fig. 9. The CED curve on 300W common set.

It contains 3837 images from the common standing datasets: AFW [4], LEN [38], LFPW [39] with annotation of 68 landmarks. To train our proposed model, we divided the 300W dataset into two parts one for training and the other for testing purposes. For training purposes, we used 3148 images, rest 689 images used for testing. We divided the testing samples into three further subgroups: (a) Common set, having 554 images (330 images of HELEN and 224 images of LFPW dataset); (b) Challenging set, having 135 images taken from IBUG dataset; (c) Full set having, all 689 images of testing part.

- COFW is a very famous publically available dataset having a total of 1852 images (1345 images are for training purposes, and the rest of 507 are for testing purposes). We used COFW to measure the performance of our proposed model. So, we just used its testing part. We used the re-annotation of [40] (68 landmarks) because originally annotated with 29 subcategories. It is very famous due to its hardness as it has diverse variations of occlusion, expressions, pose, and shape.
- AFLW is another publically available dataset having diverse 21,997 images, with 25,993 faces. Images are obtained from Flicker and have various variations in environment, age, pose, and expression. It is also annotated initially with 21 landmarks but re-annotated with 68 landmarks for FLD purposes.
- 300VW is a publically available dataset with 114 videos. All Videos are extracted into corresponding frames and annotated with 68 landmarks. We divided the 300VW dataset into two parts training and testing. For training purposes, we used 50 videos, rest 61 used for testing. We the testing samples into three further subcategories.

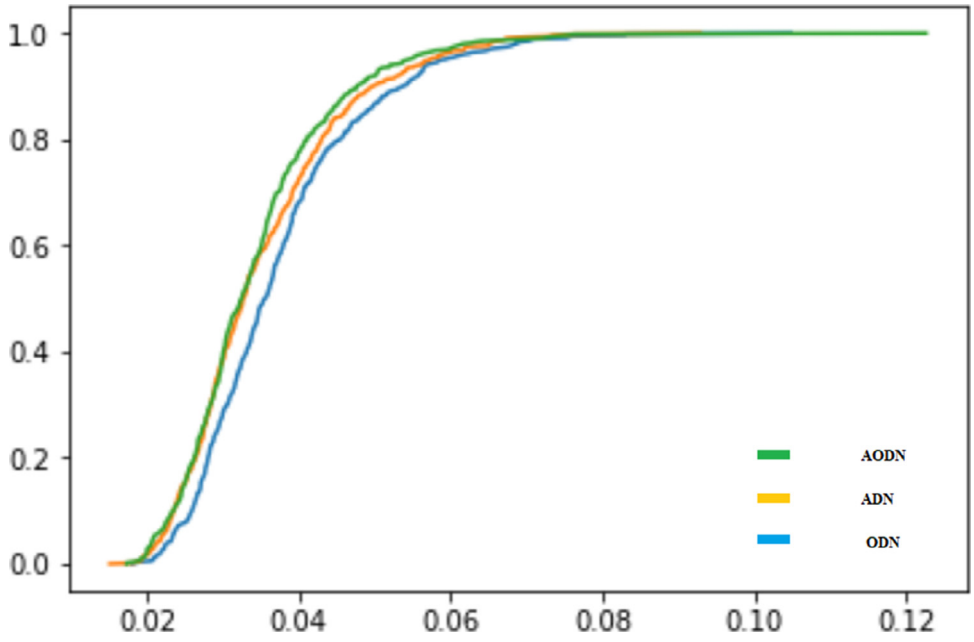

Fig. 10. The CED curve on 300W Full set.

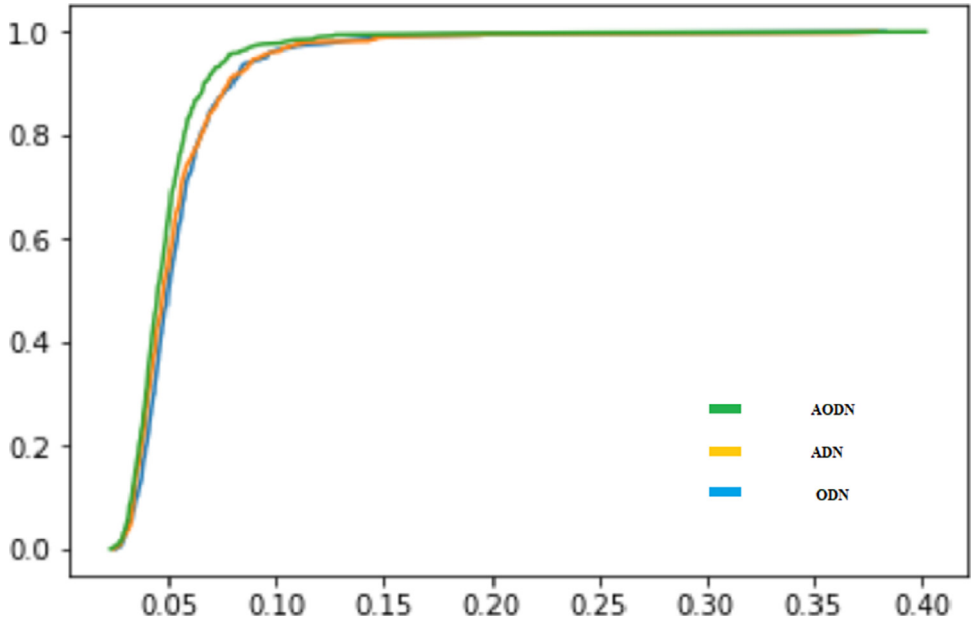

Fig. 11. The CED curve on COFW dataset.

- Menpo dataset consists of 5658 semi-frontal and 1906 profile facial images for training. For testing purposes, it consists of 5335 frontal and 1946 profile facial images. The training set is publically available, but testing data is not publically released yet. Profile facial images are annotated with 39 profile landmarks, and 68 landmarks are used for near-frontal faces. We use face images, annotated with 68 landmarks for our training. Due to the unavailability of menpo test data, we used other publically available datasets for testing purposes.

5.2. Evaluation metric and implementation details

We adopted two evaluation methods to evaluate the performance of AODN: the Cumulative Error Distribution (CED) curve, and Normalized Root Mean Squared Error (NRMSE). The NRMSE can be illustrated as:

$$NRMSE = \frac{1}{N}\sum_{i=1}^{N}\frac{\left\|\tilde{S}_i - S_i\right\|_2}{L\Omega_i} \tag{15}$$

where L represents the number of landmarks, and Ω is inter-ocular distance. In our case, Ω is the width of the bounding box of the AFLW dataset. We used different settings of parameters to measure the effectiveness, such as reduction ration r = 8, 16, 32, 64. After detailed analysis, we found r = 8 is best in our case to have better results. We also tried different combinations of CA and SA and found better performance sequentially. We used all other required parameters the same as used in ODN.

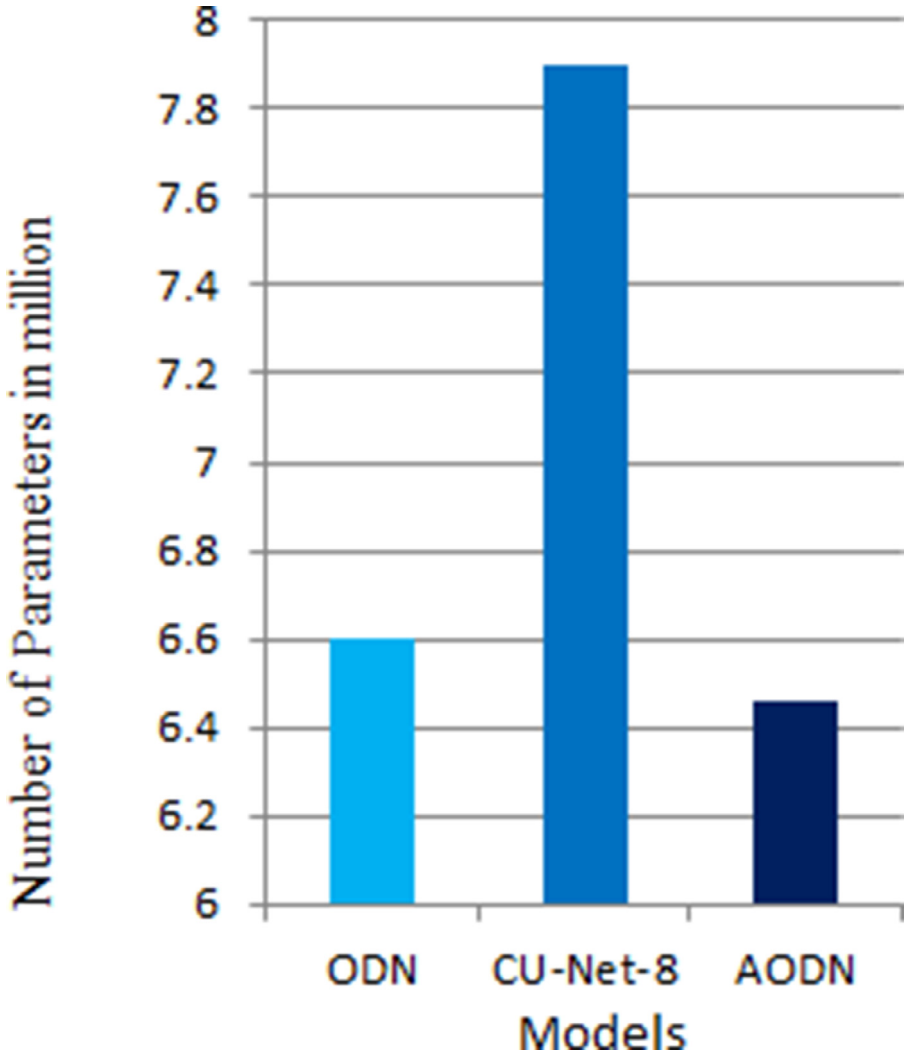

Fig. 12. The comparison of number of total parameters in millions.

5.2.1. Empirical analysis under normal circumstances

To evaluate our proposed method under normal circumstances, we used two benchmark subsets of 300W (common set and Fullset). Both subsets are benchmark datasets and have very fewer variations in illumination, pose, and occlusion. Table 1, consists of comparison results in terms of NRMSE $\left(\times 10^{-2}\right)$. We compared our results with the current best models. We can see the results for the common set; for ODN, it is 3.56 AND is 3.52. AODN reduced error from 3.56 to 3.27 for ODN and 3.52 to 3.27 for ADN, respectively. For AODN+, it is 3.14, which proves the significant change of results for the common set. Fig. 9 is about the CED curve against the common set. We can see in the CED curve AODN has significant improvement over other current state-of-the-art methods. The same as the common set, we analyzed the proposed model for the Fullset. Substantial change can be seen in Table 1, for the Fullset also. Fig. 10 is about the CED curve for the Fullset and has vigilant improvement in results in comparison to other given methods.

5.2.2. Empirical analysis for robustness against occlusion

To deal with occlusion, the first step and more important step is to model occlusion more precisely. It is hard to deal with occluded faces than regular faces. The methods perform very well to identify facial landmarks on normal facial images, decrease accuracy when to deal with occluded faces. We performed several experiments to check the robustness of AODN against occlusion. We used two diverse benchmark datasets: the challenging set of 300W and COFW to measure the effectiveness of AODN against occlusion.

The comparison results for challenging set are given in Table 2. We compare our results with the current state-of-the-art methods. It can be easily observed through the comparison table, AODN performs better and improves $\left(\times 10^{-2}\right)$ from 6.60 to 6.38, which is a significant improvement for any model. Furthermore, for AODN+, it is 5.91, which also indicates a substantial increase after training on a large scale dataset. The results of the COFW are in Fig. 7. The results of AODN for COFW are also awe-inspiring and have a significant improvement. Fig. 8 and Fig. 11 are about the CED curve of challenging set and COFW, respectively. Significant improvement in both CED curves can also be seen.

5.2.3. Empirical analysis for robustness against poses

Experimental results proved the robustness of our proposed model against normal circumstances, occlusion, Videos, as well as different poses. To further verify the generalization of our proposed method against poses, we carry out experiments on the AFLW dataset. Table 4 shows the results of our proposed model on both AFLW-Full and AFLW-Frontal. AFLW-Full dataset contains a large number of face images with challenging pose variations, which are nearly three times the number of AFLW-Frontal. Therefore, the performance on AFLW-Full shows that our method is capable of handling arbitrary head poses.

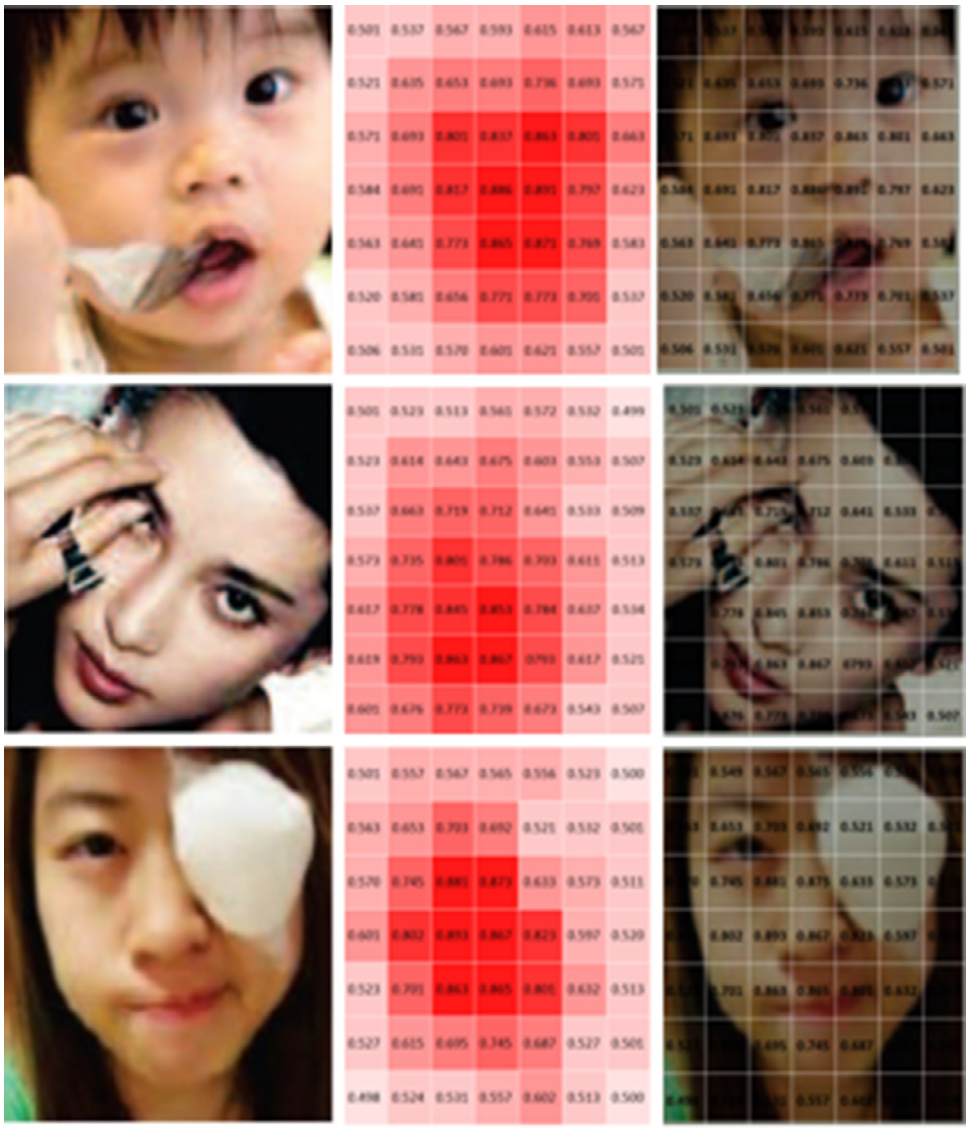

Fig. 13. The visualization of some post-attention face images along with probability map from COFW dataset. The original images are in left column, while the visualization of probability map, and post-attention results are in middle column, and right column respectively.

5.2.4. Empirical analysis for videos

To measure the effectiveness of our model, we use the 300VW dataset, which is a benchmark dataset for FLD on video. We used the testing part or 300VW, as already mentioned. For the training of our model, we used 300W for AODN and menpo for AODN+, respectively. Experimental results of all three categories in comparison to other current state-of-the-art methods are shown in Table 3. It can be easily observed that our proposed method outer performs in comparison with other methods. For Cat. 1, it improves performance from 4.75 to 4.69 for AODN, and 4.52 for AODN+. Same as Cat. 1, for Cat. 2, and Cat. 3, the improvement in performance on video dataset is significant, which proves the significance of our proposed method over other methods against all three categories.

5.3. Ablation analysis

As already discussed, AODN consists of three modules: the first one is a geometry-aware module (GM), the second module named attention module (AM) and the third is the low-rank learning module (LM). Here we perform the ablation analysis to validate the effectiveness of each module on challenging datasets. The geometry-aware module that exploits the matrix outer product to capture facial geometric relationships among different components.

Our proposed attention module simulates the mechanism of the human visual system to get rich feature representation. Furthermore, attention tells the network about the specific area that needs to be focused on. For the sake of simplification, we also use chanel-wise attention to guide the network 'what' is meaningful in a given image, and spatial attention guide the network 'where' to focus. The objective behind this attempt is to ensure the sensitivity of network to informative features. Based on the baseline ResNet-18 and AODN, we analyze the whole model and proposed changes. Table 5 shows the importance of each module as well as the robustness of our model on the challenging set of 300W. We evaluated all modules of our model as well as a different combination. We also performed our experiments against different reduction ratios, Element Wise Addition (EWA) as well as different combinations of attention. From Table 5, it can be easily observed that our proposed model consists of the best combination of various modules. Fig. 12 shows the comparison based on the number of network parameters in millions between AODN, ODN, and CU-Net-8 [45]. We just selected the most recent methods for

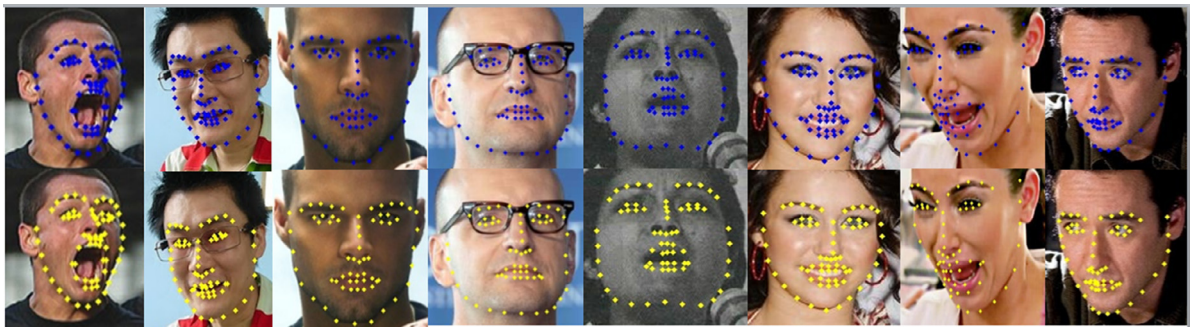

Fig. 14. Qualitative detection results of the proposed approach for some sample faces from 300W full dataset. The ground-truth landmarks are marked in blue color (top row), while our predicted landmarks are in yellow color (bottom row). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

comparison purposes. After implementing the proposed changes, AODN has a significant difference in the number of network parameters; it reduces 6.6 million to 6.46 million. This is also helpful in minimizing computation time and cost, especially in the case of scalable computing. In Fig. 13 we displayed the probability map, and visualization of some post-attention face images from COFW dataset. Fig. 14 is about the Qualitative detection results of the proposed approach for some sample faces from the 300W full dataset. The ground-truth landmarks are marked in white color (top row), while our predicted landmarks are in yellow color (bottom row).

6. Conclusion and future work

This paper has demonstrated attentive occlusion-adaptive deep network as another way to deal with FLD problems. To be very specific, we introduced channel-wise attention and spatial attention in our already established occlusion-adaptive deep network model to improve its performance. To enhance the representation of interest, model occlusion, and handle spatial distortion, we incorporated channel attention and spatial attention. The objective of channel attention is to focus on 'what' is informative in a given input facial image, and spatial attention guides the network 'where' to focus. The whole framework was tested on various benchmark datasets with different settings and compared against many state-of-the-art methods. Results proved that our proposed approach detects facial landmarks more accurately than existing methods. Taking advantage of our model's robustness, the implementation for facial expression recognition is in our next research plan. At last, we also intend to implement a parallel computing version for more efficient and fast processing.

Declaration of Competing Interest

With the submission of this manuscript, I would like to undertake that:

- All authors of this research paper have directly participated in the planning, execution, or analysis of this study;
- All authors of this paper have read and approved the final version submitted;
- The contents of this manuscript will not be copyrighted.
- The contents of this manuscript have not been copyrighted or published previously;
- The submitted manuscript is a Full-Length Research Article.

Acknowledgements

This work is supported by Ministry of Science and Technology China (MOST) Major Program on New Generation of Artificial Intelligence 2030 No. 2018AAA0102200. This work is also supported by Natural Science Foundation China (NSFC) Major Project No. 61827814 and Shenzhen Innovation Council of Science and Technology, China Project No. JCY20190808153619413. This work is also supported by the National Engineering Laboratory for Big Data System Computing Technology, China.

References

[1] I. Kemelmacher-Shlizerman, R. Basri, 3D face reconstruction from a single image using a single reference face shape, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2) (2010) 394–405.
[2] Y. Wu, Q. Ji, Facial landmark detection: a literature survey, Int. J. Comput. Vis. 127 (2) (2019) 115–142.
[3] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active shape models-their training and application, Comput. Vision Image Understanding 61 (1) (1995) 38–59.
[4] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2879–2886.
[5] G. Tzimiropoulos, J. Alabort-i Medina, S. Zafeiriou, M. Pantic, Generic active appearance models revisited, in: Asian Conference on Computer Vision, Springer, 2012, pp. 650–663.
[6] A. Asthana, S. Zafeiriou, S. Cheng, M. Pantic, Robust discriminative response map fitting with constrained local models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3444–3451.
[7] M. Zhu, D. Shi, J. Gao, Branched convolutional neural networks incorporated with jacobian deep regression for facial landmark detection, Neural Netw. (2019).
[8] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, Springer, 2014, pp. 818–833.
[9] Y. Wu, Q. Ji, Robust facial landmark detection under significant head poses and occlusion, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3658–3666.
[10] J. Xing, Z. Niu, J. Huang, W. Hu, X. Zhou, S. Yan, Towards robust and accurate multi-view and partially-occluded face alignment, IEEE Trans. Pattern Anal. Mach. Intell 40 (4) (2017) 987–1001.
[11] Q. Liu, J. Deng, J. Yang, G. Liu, D. Tao, Adaptive cascade regression model for robust face alignment, IEEE Trans. Image Process. 26 (2) (2016) 797–807.
[12] J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou, RetinaFace: single-shot multi--level face localisation in the wild, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5203–5212.
[13] M. Zhu, D. Shi, M. Zheng, M. Sadiq, Robust facial landmark detection via occlusion-adaptive deep networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3486–3496.
[14] X.P. Burgos-Artizzu, P. Perona, P. Dollár, Robust face landmark estimation under occlusion, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1513–1520.
[15] H. Li, Y. Li, J. Xing, H. Dong, Spatial alignment network for facial landmark localization, World Wide Web 22 (4) (2019) 1481–1498.
[16] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, T.-S. Chua, SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5659–5667.
[17] S. Woo, J. Park, J.-Y. Lee, I. So Kweon, CBAM: convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
[18] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[19] A. Bringmann, S. Syrbe, K. Görner, J. Kacza, M. Francke, P. Wiedemann, A. Reichenbach, The primate fovea: structure, function and development, Prog. Retin. Eye Res. 66 (2018) 49–84.
[20] A.V. Tschulakow, T. Oltrup, T. Bende, S. Schmelzle, U. Schraermeyer, The anatomy of the foveola reinvestigated, PeerJ 6 (2018) e4482.

[21] L. Yue, X. Miao, P. Wang, B. Zhang, X. Zhen, X. Cao, Attentional alignment networks, in: BMVC, vol. 2, 2018, p. 7.
[22] Z. Shao, Z. Liu, J. Cai, L. Ma, Deep adaptive attention for joint facial action unit detection and face alignment, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 705–720.
[23] J. Park, S. Woo, J.-Y. Lee, I.S. Kweon, BAM: bottleneck attention module, arXiv preprint arXiv:1807.06514 (2018).
[24] M. Sadiq, D. Shi, M. Guo, X. Cheng, Facial landmark detection via attention-adaptive deep network, IEEE Access 7 (2019) 181041–181050.
[25] F. Nie, H. Huang, C. Ding, Low-rank matrix recovery via efficient schatten p-norm minimization, in: Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012.
[26] G.H. Golub, C. Reinsch, Singular value decomposition and least squares solutions, in: Linear Algebra, Springer, 1971, pp. 134–151.
[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.
[28] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, 300 Faces in-the-wild challenge: the first facial landmark localization challenge, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2013, pp. 397–403.
[29] M. Koestinger, P. Wohlhart, P.M. Roth, H. Bischof, Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization, in: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), IEEE, 2011, pp. 2144–2151.
[30] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, J. Shen, The menpo facial landmark localisation challenge: a step towards the solution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 170–179.
[31] G. Tzimiropoulos, Project-out cascaded regression with an application to face alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3659–3667.
[32] A. Kumar, R. Chellappa, Disentangling 3D pose in a dendritic CNN for unconstrained 2D face alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 430–439.
[33] S. Honari, P. Molchanov, S. Tyree, P. Vincent, C. Pal, J. Kautz, Improving landmark localization with semi-supervised learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1546–1555.
[34] X. Dong, Y. Yan, W. Ouyang, Y. Yang, Style aggregated network for facial landmark detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 379–388.
[35] P. Gao, K. Lu, J. Xue, L. Shao, J. Lyu, A coarse-to-fine facial landmark detection method based on self-attention mechanism, IEEE Trans. Multimedia (2020).
[36] B. Browatzki, C. Wallraven, 3fabRec: fast few-shot face alignment by reconstruction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6110–6120.
[37] P. Chandran, D. Bradley, M. Gross, T. Beeler, Attention-driven cropping for very high resolution facial landmark detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5861–5870.
[38] V. Le, J. Brandt, Z. Lin, L. Bourdev, T.S. Huang, Interactive facial feature localization, in: European Conference on Computer Vision, Springer, 2012, pp. 679–692.
[39] P.N. Belhumeur, D.W. Jacobs, D.J. Kriegman, N. Kumar, Localizing parts of faces using a consensus of exemplars, IEEE Trans. Pattern Anal. Mach. Intell. 35 (12) (2013) 2930–2940.
[40] G. Ghiasi, C.C. Fowlkes, Occlusion coherence: localizing occluded faces with a hierarchical deformable part model, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2385–2392.
[41] X. Miao, X. Zhen, X. Liu, C. Deng, V. Athitsos, H. Huang, Direct shape regression networks for end-to-end face alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5040–5049.
[42] X. Dong, S.-I. Yu, X. Weng, S.-E. Wei, Y. Yang, Y. Sheikh, Supervision-by-registration: an unsupervised approach to improve the precision of facial landmark detectors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 360–368.
[43] H. Liu, J. Lu, J. Feng, J. Zhou, Two-stream transformer networks for video-based face alignment, IEEE Trans. Pattern Anal. Mach. Intell. 40 (11) (2017) 2546–2554.
[44] R. Hannane, A. Elboushaki, K. Afdel, A divide-and-conquer strategy for facial landmark detection using dual-task CNN architecture, Pattern Recognit. (2020) 107504.
[45] Z. Tang, X. Peng, K. Li, D.N. Metaxas, Towards efficient U-Nets: a coupled and quantized approach, IEEE Trans. Pattern Anal. Mach. Intell. (2019).

Mr. Sadiq received his MS(CS) degree from Riphah International University Pakistan in 2015. Mr. Sadiq is currently PhD Scholar in College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China from 2017. His research interests are Artificial Intelligence, Cloud Computing, Cloud Security, Computer Vision, etc. Mr. Sadiq have several publications in the last few years.

Daming Shi received the PhD degree in mechanical engineering from Harbin Institute of Technology, China, in 1997, and the PhD degree in computer science from University of Southampton, United Kingdom, in 2002. Before he joined Shenzhen University as a Distinguished Professor in 2016, he had been serving as a Reader / Professor at Middlesex University, UK, since 2010, and Assistant Professor at Nanyang Technological University, Singapore, during 2002–2009. He has also held an appointment of Adjunct Professor at Harbin Institute of Technology. Prof. Shi chaired the technical committee on Intelligent Internet System, IEEE SMC Society from 2005 to 2010. His current research interests include machine learning, image processing, and computer vision. He has published one book and over 150 academic papers, which appear in reputable journals such as IEEE Transactions on Pattern Analysis and Machine Intelligence and IEEE Transactions on Image Processing, etc. He has completed quite a number of projects funded by national-level research councils, such as Agency of Science and Technology Research (A*STAR) Singapore, Natural Science Foundation Council (NSFC) China, and European Council FP7
Email: dshi@szu.edu.cn