IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 22, 2019, accepted October 29, 2019, date of publication November 22, 2019,
date of current version December 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2955156

Facial Landmark Detection via Attention-Adaptive Deep Network

MUHAMMAD SADIQ[1], DAMING SHI[1], (Senior Member, IEEE), MEIQIN GUO[1], AND XIAOCHUN CHENG[2]

[1]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
[2]Faculty of Science and Technology, Middlesex University, London NW4 4BT, U.K.

Corresponding author: Daming Shi (dshi@szu.edu.cn)

This work was supported by Natural Science Foundation China (NSFC) Major Project No. 61827814 and Shenzhen Science and Technology Innovation Commission (SZSTI) Project No. JCYJ20170302153752613. It is also supported by Ministry of Science and Technology China (MOST) Major Program on New Generation of Artificial Intelligence 2030 No. 2018AAA0102202. This work was supported by the National Engineering Laboratory for Big Data System Computing Technology, China.

ABSTRACT Facial landmark detection is a key component of the face recognition pipeline as well as facial attribute analysis and face verification. Recently convolutional neural network-based face alignment methods have achieved significant improvement, but occlusion is still a major source of a hurdle to achieve good accuracy. In this paper, we introduce the attentioned distillation module in our previous work Occlusion-adaptive Deep Network (ODN) model, to improve performance. In this model, the occlusion probability of each position in high-level features are inferred by a distillation module. It can be learnt automatically in the process of estimating the relationship between facial appearance and facial shape. The occlusion probability serves as the adaptive weight on high-level features to reduce the impact of occlusion and obtain clean feature representation. Nevertheless, the clean feature representation cannot represent the holistic face due to the missing semantic features. To obtain exhaustive and complete feature representation, it is vital that we leverage a low-rank learning module to recover lost features. Considering that facial geometric characteristics are conducive to the low-rank module to recover lost features, the role of the geometry-aware module is, to excavate geometric relationships between different facial components. The role of attentioned distillation module is, to get rich feature representation and model occlusion. To improve feature representation, we used channel-wise attention and spatial attention. Experimental results show that our method performs better than existing methods.

INDEX TERMS Facial landmarks, channel attention, spatial attention, deep learning, scalable computing.

I. INTRODUCTION

Facial points are predefined landmark points on a face graph, mainly located around the common facial components, e.g. nose, chin, mouth, ear and eyes. These points also can be centred at common facial components. Facial analysis tasks can differ in numbers and types, in terms of number of needed facial points, and use of these facial points. For face alignment mostly localizing of these facial points has been done, and it gained more attention during the last decade [1] due to its importance. The area of Facial Landmark Detection (FLD) has received much attention within the computer vision community since it is a fundamental problem. In fact, fast and accurate FLD is beneficial either directly or indirectly to many application domains [1]: Directly by indicating the relevant area where to extract the interesting features as for face recognition, facial emotion recognition, and gender recognition; Indirectly as a critical task for face alignment, facial attribute analysis, dysmorphic facial signs identification in the medical field, and even for face detection [2]. Because of the practical value associated with this topic, it has been attracting efforts from both industry and academia, which has resulted in impressive progress in recent years. In spite of the achieved promising results, accurate localization of facial landmarks in unconstrained scenarios remains an extremely challenging task, mostly when faced with viewpoint change, severe occlusion and large appearance variation, caused by change in the illumination. Complex facial

The associate editor coordinating the review of this manuscript and approving it for publication was Honghao Gao.

expression and different head poses make the task even harder.

Existing facial landmark detection methods can be categorised in 3 groups: template based methods, regression based methods, and deep learning based methods. Template-based fitting methods aim to learn a shape model during training and then fit the input pictures when testing. It can be formulated by minimizing the distance between the reconstructed images and the shape normalized testing image [3]. It builds global facial shape and facial appearance based on Principal Component Analysis (PCA), mainly the focus is to improve the fitting algorithms. Notable examples include: ''Face detection, pose estimation, and Landmark Localization'' (FPLL) [4], Active Appearance Models (AAM) [5], Discriminative Response Map Fitting (DRMF) [6], and Active Shape Models (ASM) [7]. More particularly, ASM method incorporates the distributions of corresponding anatomical points, and then parameterizes the mean shape and most likely variations of this shape across a training set. As for AAM method, PCA is firstly used to model the shape and texture separately and then integrating them together with another PCA to get a generative appearance model. Various refinements on this basic scheme of AAM have boosted the performance, and have shown to be robust with large variations of occlusions and extreme illumination conditions. In this type of model, the reconstruction error affects whole face under occlusion [8], which leads models to be unable to locate facial landmarks in hard circumstances.

Regression-based methods aim to directly learn the mapping from image appearance to landmark locations. This can be done either in one iteration without any initialization of landmark positions, which is the case of direct regression methods; or by starting from an initial guess of landmark locations (e.g. a face from training set or mean face) and perform cascaded prediction, which is the case of cascaded regression methods [3].

Recently, deep learning methods achieved prominent place to solve computer vision problems. Facial landmark detection researchers shift their attention from traditional methods to deep learning based methods. Convolutional Neural Network (CNN) models are dominant deep learning based models for facial landmark detection, and most of them follow the regression framework. Those methods can either directly predict the facial landmark locations, or combine other deep learning concepts with computer vision projection models for prediction. Comparatively, deep learning based models have gained superior performance over other models [9]. Recently, convolutional neural network base facial point estimation methods have achieved significant improvement [10]–[12]. In line with the recent continuous success of deep learning in vision, the accuracy of the FLD task has significantly improved. However, occlusion and extreme facial expression images are still an existing challenging task for CNN as well [13]. If the face is partially occluded, the localizing accuracy would be dropped significantly because occlusion probably misleads CNN for feature representation learning.

FIGURE 1. Examples of occlusion from the COFW dataset [15]. It is very hard to detect the facial landmarks when face is occluded by glasses, masks, hairs, hands and scarves, etc.

CNN works in a hierarchical manner and deals with local dependencies, which depends on size of kernel and depth of network. CNN uses the concept of padding to maintain the similarity between input and output length. For L layers of CNN with kernel size of K, the largest size of context size can be $L(k-1)$ [14]. As discussed, the convolutional process deals with just local neighbour dependencies, either in space or time; thus, dependencies at long-range can only be captured through multiple repetitions of these operations, which is time and resource taking. CNN pushes performance on basis of the network's depth, width, and cardinality.

To solve the occlusion problem, the first step is to model occlusion. It is hard to model occlusion, and specifically for facial appearance. It is very challenging because it is irregular, complex, and random, as shown in Figure. 1. To solve the occlusion problem there exist some methods in literature. Robust Cascaded Pose Regression (RCPR) [15] predicts the occlusion likelihood of relevant landmarks using a fixed occlusion prior knowledge. RCPR divides the face into different blocks, and training depends on the annotated occlusion state of landmarks in the training set, which is very time consuming for large scale datasets, e.g. 300W [16], AFLW [17], etc. Wu and Ji [18] introduced a supervised regression method that gradually updates probabilities of landmark visibility in each iteration, Liu *et al.* [19] Introduced adaptive cascade regression with adaptive exemplar-based shape model to estimate the occlusion level of each landmark. Recently Xing *et al.* [20] introduced an occlusion dictionary into the face appearance dictionary, and occlusion dictionary is learned in a data driven manner. To overcome the occlusion problem during landmark detection we already proposed (ODN) [21]. ODN consists of three modules: distillation module, geometry-aware module, and low-rank learning module. We used distillation module to model occlusion probability based on high level feature, furthermore we used low-rank learning module to recover missing feature.

Although ODN achieved better results than previous methods, we observed the results are not as per our expectation due to poor feature representation. In real-world scenario, many facial images are collected in the wild, which are affected by spatial and appearance distortion due to irregularities of

positions in a camera according to the scene, which alter the dimensions of the geometry of the scene, and degrades performance [22].

To improve representation interests and to handle spatial distortion, we employ attention mechanism into already established ODN. Attention mechanism consists of channel-wise attention and spatial attention, channel attention focuses on 'what' is meaningful in a given input facial image, and spatial attention focuses on 'where' to focus.

The remaining part of the paper is organized as follows. Section II describes the related work. Section III introduces occlusion adaptive attention based solution. Section IV elaborates on experiments and results. In section V we describe conclusion and future work.

II. RELATED WORK

The prospective behind facial landmark detection is to identify some predefined key-points on facial components. Unfortunately, occlusion is still an obstacle to achieve this task perfectly.

Attention has a vital role in human perception. The human visual system doesn't process the whole scene at once it exploits a sequence of partial glimpses and selectively focuses on specific part in order to capture better visual structure. In human the human visual system only in the foveolar visual acuity reaches 100 percent due to the largest connection of cones. Attention plays a very critical role to capture long-range dependencies. Attention calculates response for a specific location as weighted sum of features at all positions, and tells network where to focus specifically. In this section we will analyze the problem with the ODN followed by the discussion about attention for FLD.

A. OCCLUSION-ADAPTIVE DEEP NETWORK

ODN consists of three modules: distillation module, geometry-aware module, and low-rank learning module. We used the distillation module to model occlusion probability based on high-level feature, furthermore we used low-rank learning module to recover missing feature. In real-world scenario many facial images are collected in the wild, which effected by spatial and appearance variations [22] due to irregularities of positions in a camera according to the scene, which alter the dimensions of the geometry of scene, and degrades performance.

To obtain ODN we modify the last residual block of ResNet-18 [23]. Feature map from last residual block is fed into geometry-aware module, distillation module to obtain geometric information, and clean feature representation. Geometry-aware module consists of two pathway subnetworks, is similar to quadratic kernel expansion, which aims to generate high level feature representation. Objective of distillation module is, to eases the sensitive of occlusion, filter the features of occluded region, and remove irrelevant information from background. Distillation modules also consist of two sub-pathway networks. To generate the hybrid feature representation of face appearance, geometry-aware module

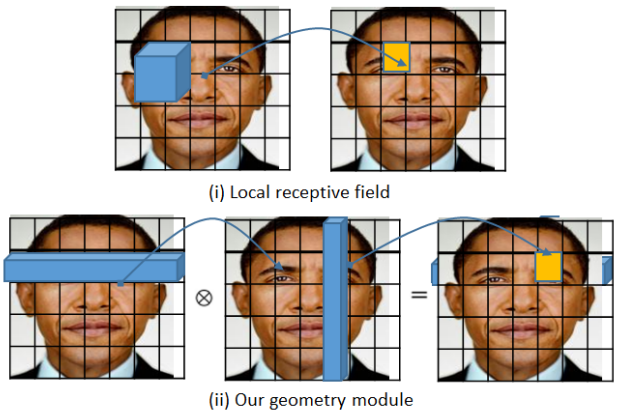

FIGURE 2. Comparison of local receptive field and our proposed geometry-aware module on capturing facial geometric relations.

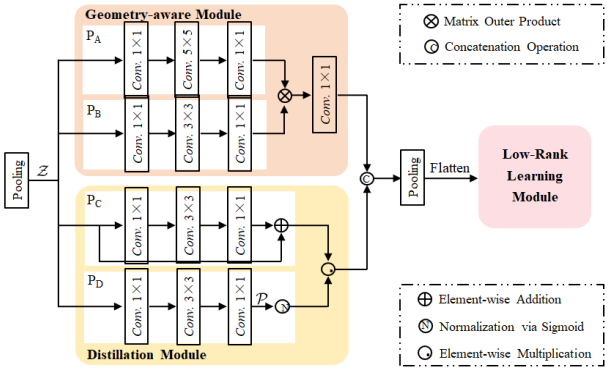

FIGURE 3. Occlusion-adaptive Deep Network.

and distillation module are concatenated together into one high-dimensional feature map. The output of geometry-aware module and distillation module are assembled and fed into low-rank learning module to recover missing features by modelling facial inter-feature correlation. In ODN feature distillation module is responsible to model occlusion and feature representation. Rich feature representation can help low rank learning module to recover missing feature in efficient way.

B. ATTENTION ADAPTIVE FACIAL ALIGNMENT

Attention has a vital role in human perception. Normally human visual system process iteratively, instead of processing the whole image at once, exploits a sequence of partial glimpses and focus on selective portions to capture better visual structure [24]–[28]. In the human visual system only in the foveolar visual acuity reaches 100 percent [29]–[33], due to the largest connection of the cones.

Recently there are several attempts to improve the performance of CNN by using attention [34]–[38]. Attention calculates response for a specific location as the weighted sum of the features at all positions, and indicates where to focus [34]. Li *et al.* [22] introduced Spatial Alignment Network (SAN) for facial landmark detection, focusing on spatial and appearance variations. SAN consists of two methods;

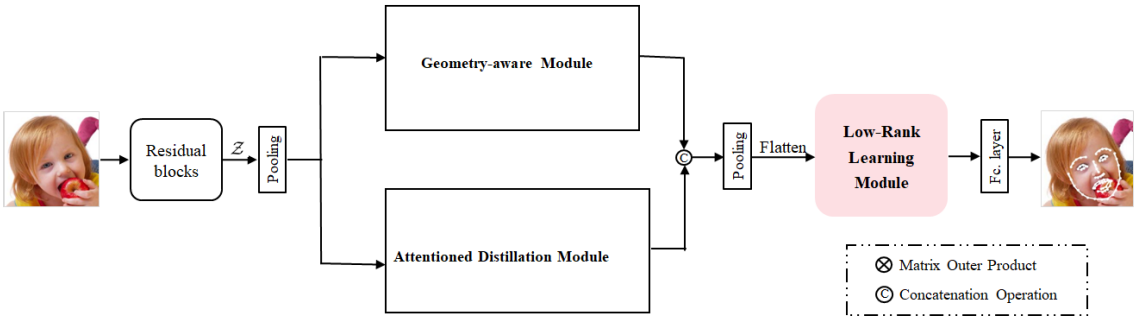

FIGURE 4. The diagram of Occlusion-adaptive Attention Deep Network Model.

the hand-crafted method, as well as learning-based method. But, the problem is, if the handcrafted method is used, the efficiency is very low and if the learning-base method is used, it is not steady. JAA-Net [39] and AAN [40] uses spatial attention to improve network performance regarding landmark detection. In JAA-Net, and AAN spatial alignment indicates 'where' to focus but still 'what' to focus is missing. Channel-wise attention and spatial attention [34], [35], [41] achieved remarkable improvement for image classification. Inspired by [34], [35], [41] we incorporated channel-wise attention and spatial attention, in already established ODN [21] to improve the performance of our model.

III. OCCLUSION ADAPTIVE ATTENTION BASED DEEP NETWORK

Most deep learning based face alignment models are using CNN. Convolutional process deals with just local, neighbour dependencies. Thus, dependencies at long-range can be captured through multiple repetitions of these operations, so it is computationally inefficient to compute long-range dependencies. Attention calculates responses for a specific location as the weighted sum of the features at all positions and leads, where to focus [34]. Channel-wise attention and spatial attention achieved remarkable improvement in performance of the CNN. We incorporated channel-wise attention and spatial attention in the ODN [21] to improve the performance of network.

To be very specific in ODN we edited the last Residual block of ResNet18 [23]. It consists of three modules: distillation module, geometry-aware module, and low-rank learning module. We replaced distillation module with attentioned distillation module by incorporating channel attention and spatial attention. We used the attentioned distillation module to model occlusion probability based on high level feature, furthermore we used low-rank learning module to recover missing feature.

The attentioned distillation module filters the feature of occluded region. The absence of some features doesn't mean, the face doesn't have that features, which could be an incorrect interpretation of the model. Large number of feature of face are co-related or co-occur. Some have symmetry, Proximity or position relation. Presence of features directs towards the presence of other features or recover missing features.

Layer name	output size	18-layers	No. of Blocks
conv1	112x112	7x7,64, stride 2	
conv2_x	56x56	3x3 max pool, stride 2	
		3x3,64	x 2
		3x3,64	
conv3_x	28x28	3x3,128	x 2
		3x3,128	
conv4_x	14x14	3x3,256	x 2
		3x3,256	
conv5_x	7x7	3x3,512	x 2
		3x3,512	
	1x1	average pool, 100-d fc, softmax	
FLOPs	1.8 x 10*9		

FIGURE 5. ResNet18 [23] Default architecture.

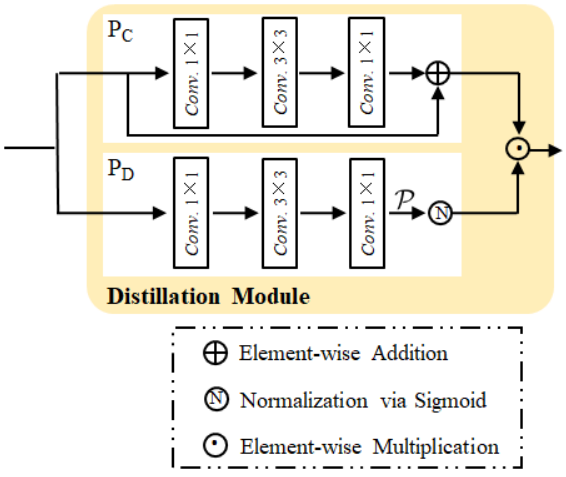

FIGURE 6. Distillation module.

In real-world scenario many facial images are collected in the wild, which effected by spatial and appearance distortion due to irregularities of positions in a camera according to the scene, which alter the dimensions of the geometry of the scene, and degrades performance.

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

Given the training set $\left\{ \left(I_i, \check{S}_i \right) \right\}$ can be learned by (1).

FIGURE 7. Self-Attention module.

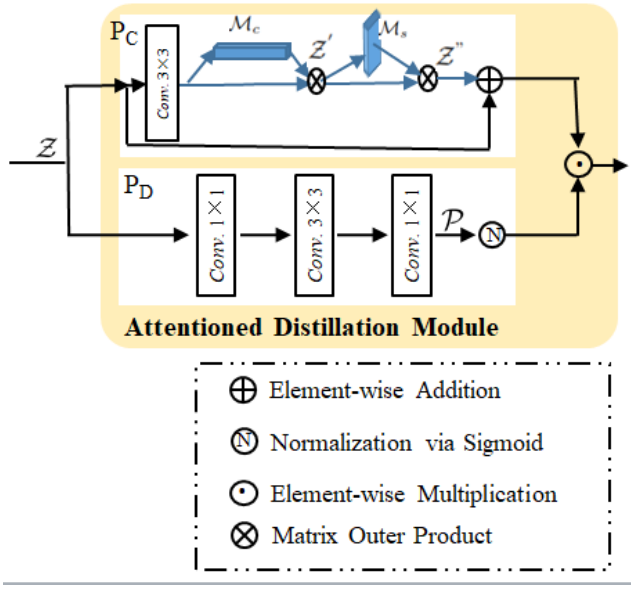

FIGURE 8. Proposed attentioned distillation module.

where $\breve{S}$, and S represents ground-truth and corresponding prediction. $\breve{S} = \{s_1, s_2, \ldots .s_L\}$ and $S = W_{fc}^T \mathcal{M}^T \mathcal{X}$. where $\mathcal{X}$ denotes the output of the geometry-aware module and the attentioned distillation module as feature map. L is the number of landmark and s denotes the facial landmark. W_{fc} denotes the parameters of fully connected layer.

We have trained our model same as the ODN in an end-to-end manner.

A. ATTENTIONED DISTILLATION MODULE

As the self-attention, in the visual context is designed to explicitly learn the relationship between one pixel and all other positions, even regions far apart, which is computationally so expensive and time-consuming.

In Figure 8. $\mathcal{Z}$ is the feature map from previous residual learning blocks. It can be easily observed, ODN needs good feature representation. Good network feature representation can help to learn missing feature more efficiently. To achieve this goal we incorporated channel-wise attention and spatial attention as proved by [34], [35], [41], that channel-wise attention and spatial attention can improve the ability of feature representation. The attentioned distillation module also consist of two pathways. Pathway-C deploys the attention mechanism to improve feature representation.

$$\mathcal{Z} = \mathcal{Z}'' + \mathcal{F} \tag{2}$$

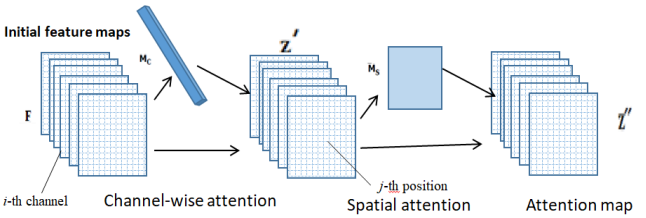

FIGURE 9. Attention mechanism.

where $\mathcal{Z}$ is a refined feature map after combining attention map and residual map. $\mathcal{F}$ is feature map from previous residual block and $\mathcal{Z}''$ is feature map from attention process.

B. CHANNEL-WISE ATTENTION AND SPATIAL ATTENTION

Normally in network engineering, to improve rich feature representation researchers increase the depth of the network. Recently [35], [41] proved that feature representation can be improved through attention mechanism. Attention not just improve feature representation, It also guides the network about, area to be focused. So in this work we will use channel-wise attention and spatial attention. To be specific channel wise attention tell the network "what" to be focused and spatial attention guides "where" to be focused. In our case spatial attention guides the network about the location to predict landmarks and channel-wise attention helps to select semantic features. Our goal is to ensure that the network is able to increase its sensitivity to informative features. Channel attention map can be produced easily by exploiting inter-channel relationship to extract semantic features for specific facial points.

CNN generally extracts features of images. An input image $W \times H \times 3$, when passes through a convolutional layer having C channels, filters scan the input image and generate output as $\acute{W} \times \acute{H} \times C$ feature map, which will be input of the next convolutional layer. In CNN, normally filters perform as pattern detector i.e. lower-level filters detect low level visual patterns like corners and edges and the high-level filter detects semantic patterns like parts, objects etc. CNN extracts image features by stacking of layers through a hierarchy of visual abstraction. So in simple words we can say image features of CNN are spatial, channel-wise and multi-layer. In our case in context of facial point, channel-wise attention can be viewed as the process of selecting semantic attributes on demand. For example when we wants to predict facial points of the left eye, our channel-wise attention will assign more weights on channel-wise feature map generated by filters.

1) CHANNEL-WISE ATTENTION

The channel attention map has been produced by exploiting the inter-channel relationship of the features. As it is a common perception that each channel of a feature map is a feature detector. As already discussed, channel attention focuses on 'what' is meaningful in the given input image. We squeeze the spatial dimension of the input feature map to compute the channel attention efficiently. In ODN we

used average-pooling but here, used both max-pooling and average-pooling. The primary purpose of max-pooling is to gather more valuable information about the distinctive features of the object to refine channel-wise attention.

To compute channel attention we first aggregated spatial information of a feature map by using both average-pooling and max-pooling operations. Later we generated two different spatial context descriptors: $\mathcal{Z}^c_{Avg}$ and $\mathcal{Z}^c_{Max}$, which denote average-pooled features and max-pooled features respectively. A shared network with these both descriptor will produce our channel attention map $\mathcal{M}_c \in \mathcal{R}^{1\times1\times C}$. The shared network is composed of multi-layer perceptron (MLP) with one hidden layer. To reduce parameter overhead, the hidden activation size is set to $\mathcal{R}^{1\times1\times\frac{C}{r}}$, where r is the reduction ratio. After the shared network is applied to each descriptor, we merge the output feature vectors using element-wise summation.

$$\mathcal{Z}' = \mathcal{M}_c \bigotimes \mathcal{F} \tag{3}$$

$\mathcal{M}_c, \mathcal{F}$ represents the channel wise attention and feature map from previous residual blocks respectively.

2) SPATIAL ATTENTION

The spatial attention map is generated by utilizing the inter-spatial relationship between features. As we already discussed the basic purpose of spatial attention is to guide the network "where" to focus, and it is complementary to the channel attention. Spatial attention can be computed by applying pooling along with channel axis, so we applied average pooling and max pooling along with channel axis and then concatenated them to generate the feature descriptor, and applied convolutional layer to generate a spatial attention map $\mathcal{M}_s\{\mathcal{Z}\} \in \mathcal{R}^{H\times W}$ which encodes where to emphasize. The channel information of a feature map has been aggregated by using two pooling operations, generating two 2D maps: $\mathcal{Z}^s_{avg} \in \mathcal{R}^{H\times W\times1}$ and $\mathcal{Z}^s_{max} \in \mathcal{R}^{H\times W\times1}$.

$$\mathcal{Z}'' = \mathcal{M}_s \bigotimes \mathcal{Z}' \tag{4}$$

$\mathcal{M}_s$ represents spatial attention, and $\mathcal{Z}'$ represents the feature map driven by channel wise attention.

IV. EXPERIMENTS

To test the effectiveness of the proposed model, we analysed our model for several benchmark datasets against normal circumstances, occlusion as well as against various poses. As per ODN we cropped and resized all images as (224×224) and exploit the scale, rotation, flip operation and translation to conduct data augmentation for the training set. All models are pre-trained on the ImageNet dataset [42].

A. DATASETS

To conduct the experiments we used following diverse benchmark datasets: 300W [16], AFLW [17], COFW [15], 300VW [16], [43] as all datasets are publicly available and considered as benchmark dataset. We compare our results with state-of-the-art methods [12], [40], [44], [45].

- 300W (300 Faces In-the-Wild Challenge) dataset [35] is a widely used database for evaluating near-frontal face alignment. Each face is annotated with 68 landmarks. To keep consistent with previous work, the dataset is partitioned as follows: The 300W training set with 3148 training images from AFW [4], LFPW [46] and HELEN [47]; the common testing subset with 554 test images from LFPW and HELEN; and the challenging testing subset with 135 test images from IBUG. The full testing set of 300W is the union of both common and challenging subsets. We use 3,148 images for training and 689 for testing as samples. We split testing samples in 3 subsets: (i) Challenging set (135 images from IBUG); (ii) Common set consist of 554 images (224 images from LFPW and 330 from HELEN test set); the fullset consist of 689 images (containing all testing images).
- COFW (Caltech Occluded Faces in the Wild) dataset is designed to benchmark face landmark algorithms in realistic conditions, which includes heavy occlusions and large shape variations. All images have large variation in pose, shape, occlusion and expression. Originally COFW is annotated with 29 landmarks, [48] re-annotated with 68 landmarks for landmark detection.
- AFLW (Annotated Facial Landmarks in the Wild) dataset is a large-scale, multi-view, real-world face database with annotated facial landmarks. This database includes a total of 24386 face images gathered from Flickr, having diverse variations of face appearances like: pose, age, gender, expression as well as different environmental conditions. It is also a publicly available database annotated with 21 landmarks.
- 300VW dataset contains 114 videos. Each video is one minute long and annotated with 68 facial points. 300VW includes 50 videos for training, and the remaining 64 testing videos divide further into three different categories.

B. EVALUATION METRIC AND IMPLEMENTATION DETAILS

To evaluate our proposed method we used Normalized Root Mean Squared Error (NRMSE) and CED curve. NRMSE is defined as.

$$NRMSE = \frac{1}{N}\sum_{i=1}^{N}\frac{\left\|\breve{S}_i - S_i\right\|_2}{L\Omega_i} \tag{5}$$

where L is the number of landmarks and Ω in inter-ocular distance. To be very specific Ω represents the width of bounding box for AFLW dataset. we conducted experiments with various reduction ratios like: r = 8, 16, 32 and 64. We found that r = 8 is best for our experiments. We also tried different combination of spatial attention and channel attention, and we found better performance in sequence as channel attention

TABLE 1. Comparison of NRMSE $\left(\times 10^{-2}\right)$ results on common set and full set of 300W.

Method	Year	Common Set	Fullset
DRMF [6]	2013	6.65	9.22
CFAN [49]	2014	5.50	-
CFSS [50]	2015	4.73	5.99
DR [10]	2016	4.51	6.30
DCRFA [11]	2016	4.19	5.02
RDR [51]	2017	5.05	5.80
SCNN [52]	2017	5.43	6.30
TSR [44]	2017	4.36	4.99
Seq_MT [53]	2018	4.20	4.90
PCD-CNN [12]	2018	3.67	4.44
ODN [21]	2019	3.56	4.17
ADN	**2019**	**3.50**	**4.14**

TABLE 2. Comparison of NRMSE $\left(\times 10^{-2}\right)$ results on Challenging set of 300W.

Method	Year	Challeenging set
CMD [54]	2013	19.54
CPR-RPP [55]	2015	11.57
DR [10]	2016	13.80
LBF [56]	2016	11.98
RDR [51]	2017	8.95
DVLN [57]	2017	7.62
TSR [44]	2017	7.56
DSRN [58]	2018	9.68
SBR [59]	2018	7.58
SAN [45]	2018	7.55
ODN [21]	2019	6.67
ADN	**2019**	**6.60**

and spatial attention. All other parameters are manually tuned as per ODN.

1) EVALUATION UNDER NORMAL CIRCUMSTANCES

In this subsection we will, evaluate our method on faces, under normal circumstances. We used two subsets of 300W (full set and common set) as the test datasets. The reason to choose these datasets is, because of fewer changes under the poses, occlusion, and illumination. Table 1 shows the experimental results in comparison to existing benchmark methods in terms of NRMSE $\left(\times 10^{-2}\right)$. It can easily be observed from Table 1, that our method outperforms than current state-of-the-art methods, and obtains a significant improvement on both test datasets (full-set and common-set). Hence, our results indicate that our method can accurately locate landmarks of the face under normal circumstances.

2) EVALUATION OF ROBUSTNESS AGAINST OCCLUSION

To deal with occluded faces is comparatively harder than normal faces. To the best of our knowledge, most of the

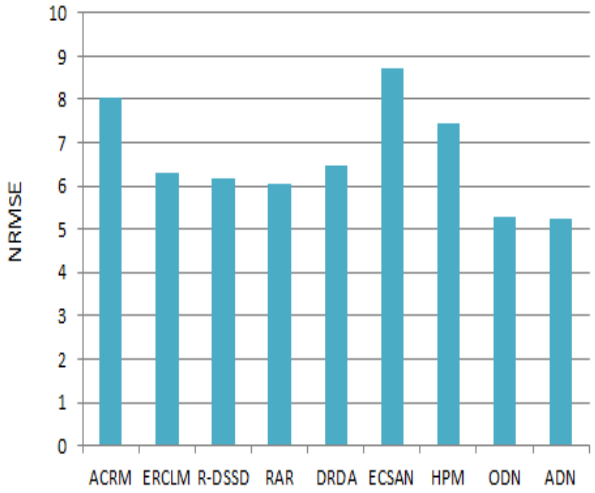

FIGURE 10. Comparison of NRMSE $\left(\times 10^{-2}\right)$ on COFW dataset.

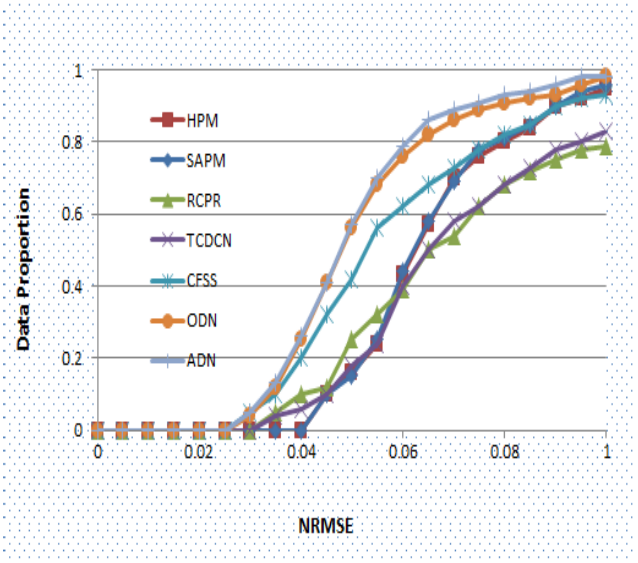

FIGURE 11. Comparison of CED curve on COFW test dataset.

state-of-the-art methods perform well under normal circumstances to predict accurate landmarks of face, but accuracy decline in case of occluded faces. Hence, to prove robustness of our method against the occlusion, we conducted experiments on two different benchmark datasets: COFW and challenging set of 300W.

As illustrated in Table 2 and Figure 10, we compared the proposed method with other current representative methods on COFW and Challenging set of 300W. The results in Table 2 show that our method boost the NRMSE value to 6.65 $\left(\times 10^{-2}\right)$, which is comparatively better then all mentioned methods on challenging set of 300W test dataset. Figure 10 shows the result on COFW test dataset. We trained our method 300W training dataset and evaluated on different datasets. From Figure 10, the improvement of result can be easily observed. Figure 11 is about CED comparison about COFW dataset. Figure 12 shows the CED comparison against challenging set of 300W dataset.

TABLE 3. Comparison of NRMSE $\left(\times 10^{-2}\right)$ results on 300VW Dataset for all 3 categories.

Method	Year	Cat. 1	Cat. 2	Cat. 3
SDM [60]	2013	7.41	6.18	13.04
TSCN [61]	2014	12.54	7.25	13.13
CFSS [50]	2015	7.68	6.42	13.67
TCDCN [62]	2015	7.66	6.77	14.98
TSTN [63]	2017	5.36	4.51	12.84
AAN [40]	2018	5.03	4.82	7.98
ADN	**2019**	**4.69**	**4.34**	**6.72**

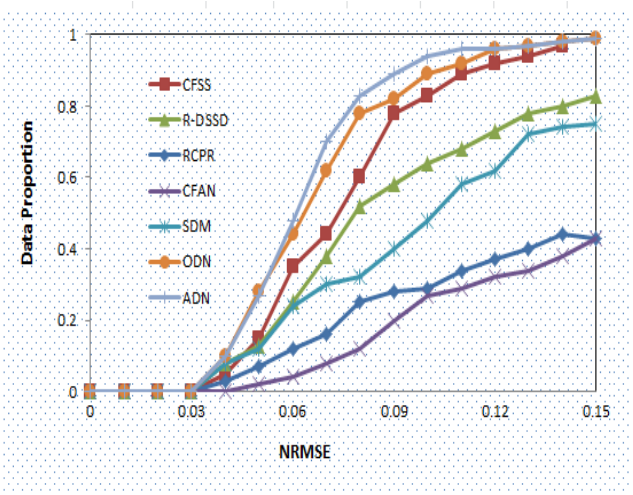

FIGURE 12. Comparison of CED curve on challenging test.

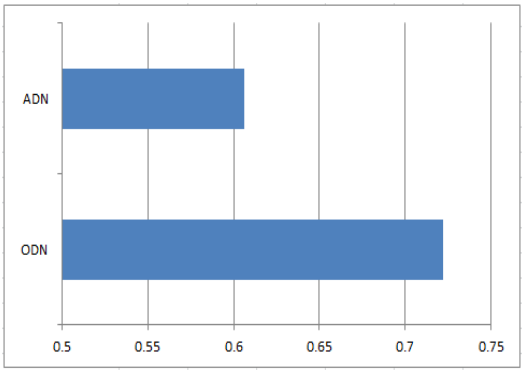

FIGURE 13. Comparison on the number of parameters in million for distillation module.

3) EVALUATION ON VIDEOS

We tested our model on 300VW dataset to asses robustness of our model. Table 3 shows the results of our model. It can be easily observed our model outperform than current state-of-the-art models against all three categories.

4) ABLATION STUDY

As already mentioned, our model is an extension of ODN. Our model consists of 3 modules: geometry-aware

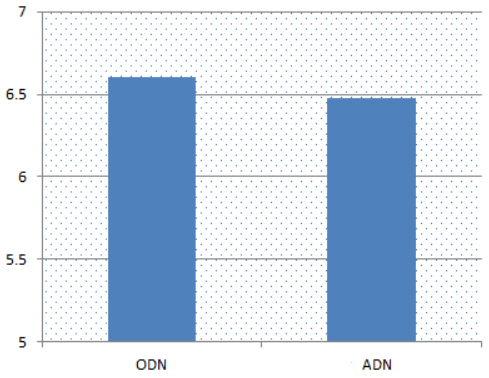

FIGURE 14. Comparison on the number of total parameters in million.

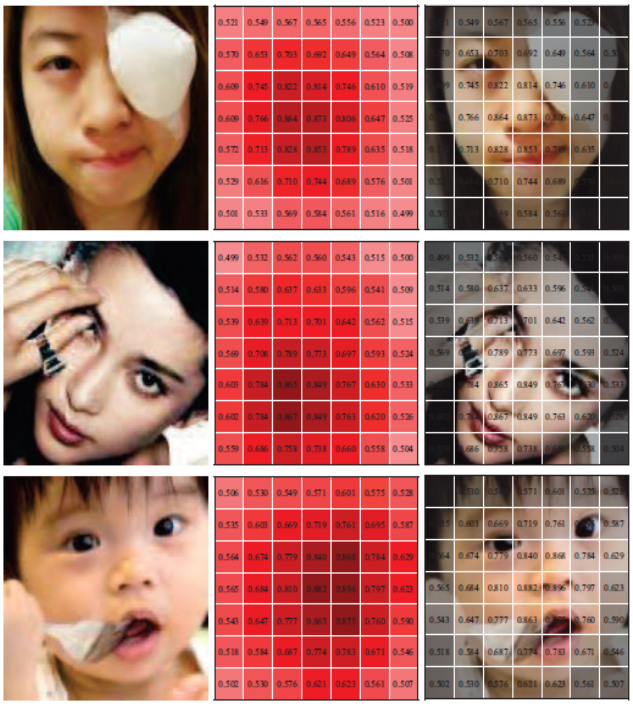

FIGURE 15. The visualization of some post-distilled face images from COFW dataset.

module (GM), attentioned distillation module (ADM) and low-rank learning module (LM). In this subsection, we carry out the validation study to validate effectiveness on challenging datasets. Based on the baseline ResNet-18 and ODN, we analyse the proposed change. Figure 13, and Figure 14 depicts the comparison on the number of network parameters in million for attentioned distillation module and overall respectively. It can be easily observed that our proposed change decreased the number of network parameters, which is time-saving as well as cost-saving. In the case of scalable data processing our model is consuming very less resources than ODN. In addition we show visualization samples from attentioned distillation module in Figure 15. In first column we can see face images and probability maps and post-distilled results of these images are illustrated in next two columns.

V. CONCLUSION AND FUTURE WORK

In this work, we present an attention-adaptive deep network to solve the occlusion problem for facial landmark detection, which is composed of three main modules: the geometry-aware module, the attentioned distillation module and the low-rank learning module. The geometry-aware module and the attentioned distillation module can capture the geometric relations of different facial components and obtain the clean feature representation, respectively. The outputs of these two modules are concatenated as the input of the low-rank learning module to recover the missing features by means of geometric information. We conducted the experiments on benchmark datasets to evaluate the performance of our proposed framework under normal circumstances, partial occlusion and extreme pose. The experimental results show that our method outperforms existing methods and achieves robustness against occlusion and various poses.

REFERENCES

[1] X. Jin and X. Tan, "Face alignment in-the-wild: A Survey," *Comput. Vis. Image Understand.*, vol. 162, pp. 1–22, Sep. 2017.

[2] M. Zhu, D. Shi, and J. Gao, "Branched convolutional neural networks incorporated with jacobian deep regression for facial landmark detection," *Neural Netw.*, vol. 118, pp. 127–139, Oct. 2019.

[3] Y. Wu and Q. Ji, "Facial landmark detection: A literature survey," *Int. J. Comput. Vis.*, vol. 127, no. 2, pp. 115–142, 2019.

[4] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2879–2886.

[5] G. Tzimiropoulos, J. Alabort-I-Medina, S. Zafeiriou, and M. Pantic, "Generic active appearance models revisited," *Asian Conference on Computer Vision*. Springer, 2012, pp. 650–663.

[6] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3444–3451.

[7] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Comput. Vis. Image Underst.*, vol. 61, no. 1, pp. 38–59, Jan. 1995.

[8] J. Zhang, M. Kan, S. Shan, and X. Chen, "Occlusion-free face alignment: Deep regression networks coupled with de-corrupt autoencoders," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3428–3437.

[9] M. Zhu, D. Shi, S. Chen, and J. Gao, "Branched convolutional neural networks for face alignment," in *Pacific Rim Conference on Multimedia*. Springer, 2018, pp. 291–302.

[10] B. Shi, X. Bai, W. Liu, and J. Wang, "Face alignment with deep regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 183–194, Jan. 2016.

[11] H. Lai, S. Xiao, Y. Pan, Z. Cui, J. Feng, C. Xu, J. Yin, and S. Yan, "Deep recurrent regression for facial landmark detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 5, pp. 1144–1157, May 2016.

[12] A. Kumar and R. Chellappa, "Disentangling 3D pose in a dendritic CNN for unconstrained 2D face alignment," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 430–439.

[13] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. ECCV*, vol. 12, 2017, pp. 40–46.

[14] G. Tang, M. Müller, A. Rios, and R. Sennrich, "Why self-attention? A targeted evaluation of neural machine translation architectures," vol. 14, no. 1, Aug. 2018, pp. 41–48, *arXiv:1808.08946*. [Online]. Available: https://arxiv.org/abs/1808.08946

[15] X. P. Burgos-Artizzu, P. Perona, and P. Dollar, "Robust face landmark estimation under occlusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1513–1520.

[16] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark Localization Challenge," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 397–403.

[17] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2144–2151.

[18] Y. Wu and Q. Ji, "Robust facial landmark detection under significant head poses and occlusion," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3658–3666.

[19] Q. Liu, J. Deng, J. Yang, G. Liu, and D. Tao, "Adaptive cascade regression model for robust face alignment," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 797–807, Feb. 2017.

[20] J. Xing, Z. Niu, J. Huang, W. Hu, X. Zhou, and S. Yan, "Towards robust and accurate multi-view and partially-occluded face alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 987–1001, Apr. 2018.

[21] M. Zhu, D. Shi, M. Zheng, and M. Sadiq, "Robust facial landmark detection via occlusion-adaptive deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3486–3496.

[22] H. Li, Y. Li, J. Xing, and H. Dong, "Spatial alignment network for facial landmark localization," *World Wide Web*, vol. 22, no. 4, pp. 1481–1498, 2018.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, vol. 4, Jun. 2016, pp. 770–778.

[24] H. Larochelle and G. E. Hinton, "Learning to combine foveal glimpses with a third-order Boltzmann machine," in *Advances in Neural Information Processing Systems*, 2010, pp. 1243–1251.

[25] M. Yin, Z. Wu, D. Shi, J. Gao, and S. Xie, "Locally adaptive sparse representation on Riemannian manifolds for robust classification," *Neurocomputing*, vol. 310, pp. 69–76, Oct. 2018.

[26] H. Gao, W. Huang, Y. Duan, X. Yang, and Q. Zou, "Research on cost-driven services composition in an uncertain environment," *J. Internet Technol.*, vol. 20, no. 3, pp. 755–769, 2019.

[27] Y. Yin, L. Chen, Y. Xu, J. Wan, H. Zhang, and Z. Mai, "QoS prediction for service recommendation with deep feature learning in edge computing environment," *Mobile Netw. Appl.*, 2019, pp. 1–11.

[28] Y. Yin, J. Xia, Y. Li, W. Xu, and L. Yu, "Group-wise itinerary planning in temporary mobile social network," *IEEE Access*, vol. 7, pp. 83682–83693, 2019.

[29] A. Bringmann, S. Syrbe, K. Görner, J. Kacza, M. Francke, P. Wiedemann, and A. Reichenbach, "The primate fovea: Structure, function and development," *Prog. Retinal Eye Res.*, vol. 66, pp. 49–84, Sep. 2018.

[30] A. V. Tschulakow, T. Oltrup, T. Bende, S. Schmelzle, and U. Schraermeyer, "The anatomy of the foveola reinvestigated," *PeerJ*, vol. 6, Mar. 2018, Art. no. e4482.

[31] Y. Yang and J. Jiang, "Hybrid sampling-based clustering ensemble with global and local constitutions," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 5, pp. 952–965, May 2016.

[32] X. Zhang, Z. Li, M. Constable, K. L. Chan, Z. Tang, and G. Tang, "Pose-based composition improvement for portrait photographs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 3, pp. 653–668, Mar. 2019.

[33] X. Wang, L. Ma, S. Kwong, and Y. Zhou, "Quaternion representation based visual saliency for stereoscopic image quality assessment," *Signal Process.*, vol. 145, pp. 202–213, Apr. 2018.

[34] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 11211. Cham, Switzerland: Springer, 2018, pp. 3–19.

[35] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "BAM: Bottleneck attention module," in *Proc. BMVC*, 2018, pp. 1–14.

[36] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[37] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. CVPR*, Jul. 2017, vol. 38, no. 1, pp. 6450–6458.

[38] J. Jiang and X. Song, "An optimized higher order CRF for automated labeling and segmentation of video objects," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 506–516, Mar. 2016.

[39] Z. Shao, Z. Liu, J. Cai, and L. Ma, "Deep adaptive attention for joint facial action unit detection and face alignment," in *Proc. ECCV*, vol. 2, Sep. 2018, pp. 725–740.

[40] L. Yue, X. Miao, P. Wang, B. Zhang, X. Zhen, and X. Cao, "Attentional alignment networks," in *Proc. BMVC*, 2018, pp. 1–14.

[41] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6298–6306.

[42] J. Deng, W. Dong, R. Socher, L.-J. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[43] G. G. Chrysos, E. Antonakos, S. Zafeiriou, and P. Snape, "Offline deformable face tracking in arbitrary videos," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 954–962.
[44] J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou, "A deep regression architecture with two-stage re-initialization for high performance facial landmark detection," in *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3691–3700.
[45] X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Style aggregated network for facial landmark detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 379–388.
[46] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2930–2940, Dec. 2013.
[47] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *European Conference on Computer Vision*. Springer, 2012, pp. 679–692.
[48] G. Ghiasi and C. C. Fowlkes, "Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1899–1906.
[49] J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment," in *European Conference on Computer Vision* Springer, 2014, pp. 1–16.
[50] S. Zhu, C. Li, C. C. Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vols. 7–12, Jun. 2015, pp. 4998–5006.
[51] S. Xiao, J. Feng, L. Liu, X. Nie, W. Wang, S. Yan, and A. Kassim, "Recurrent 3D-2D dual learning for large-pose facial landmark detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1642–1651.
[52] A. Jourabloo, M. Ye, X. Liu, and L. Ren, "Pose-invariant face alignment with a single CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3219–3228.
[53] S. Honari, P. Molchanov, S. Tyree, P. Vincent, C. Pal, and J. Kautz, "Improving landmark localization with semi-supervised learning," in *Proc. CVPR*, Jun. 2018, pp. 1546–1555.
[54] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas, "Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1944–1951.
[55] H. Yang, X. He, X. Jia, and I. Patras, "Robust face alignment under occlusion via regional predictive power estimation," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2393–2403, Aug. 2015.
[56] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment via regressing local binary features," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1233–1245, Mar. 2016.
[57] W. Wu and S. Yang, "Leveraging intra and inter-dataset variations for robust face alignment," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 2096–2105.
[58] X. Miao, X. Zhen, X. Liu, C. Deng, V. Athitsos, and H. Huang, "Direct shape regression networks for end-to-end face alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5040–5049.
[59] X. Dong, S.-I. Yu, X. Weng, S.-E. Wei, Y. Yang, and Y. Sheikh, "Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, no. 1, Jun. 2018, pp. 360–368.
[60] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 532–539.
[61] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.
[62] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning deep representation for face alignment with auxiliary attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 918–930, May 2016.
[63] H. Liu, J. Lu, J. Feng, and J. Zhou, "Two-stream transformer networks for video-based face alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2546–2554, Nov. 2017.

MUHAMMAD SADIQ received the M.S. degree in computer science from Riphah International University, Pakistan, in 2015. He is currently pursuing the Ph.D. degree with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. He has published several publications in the last few years. His research interests are artificial intelligence, cloud computing, cloud security, computer vision, and so on.

DAMING SHI (M'02–SM'04) received the Ph.D. degree in mechanical engineering from the Harbin Institute of Technology, China, and the Ph.D. degree in computer science from the University of Southampton, U.K. He has been serving as a Distinguished Professor with Shenzhen University, since 2016. Before that, he worked with Middlesex University, U.K., Kyungpook National University, South Korea, and Nanyang Technological University, Singapore. His current research interests include machine learning, image processing, pattern recognition, and neural networks.

MEIQIN GUO received the B.S. degree from the School of Nanchang University, Nanchang, China, in 2015. He is currently pursuing the M.S. degree with the Department of Software Engineering, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. His current research interests include digital image processing and deep learning.

XIAOCHUN CHENG received the bachelor's degree in computer engineering, in 1992, and the Ph.D. degree in computer science from Jilin University, in 1996. He has been a Computer Science EU Project Coordinator with Middlesex University, since 2012. He contributed for peer-reviewed 103 published journal articles, 134 conference papers, with five times best conference paper awards by the end of 2018. He is a member of the IEEE SMC Technical Committee on Enterprise Information Systems, the IEEE SMC Technical Committee on Computational Intelligence, the IEEE SMC Technical Committee on Cognitive Computing, the IEEE SMC Technical Committee on Intelligent Internet Systems, and the BCS Artificial Intelligence Specialist Group. Four papers are in the 2018 top 1% of the academic field based on a highly cited threshold for the field and publication year by data from Essential Science Indicators.