Robust Facial Landmark Detection via Occlusion-adaptive Deep Networks

Meilu Zhu, Daming Shi*, Mingjie Zheng, Muhammad Sadiq
College of Computer Science and Software Engineering, Shenzhen University, China
{zhumeilu2016,zhengmingjie}@email.szu.edu.cn; dshi@szu.edu.cn; sadiq_paec@yahoo.com

Abstract

In this paper, we present a simple and effective framework called Occlusion-adaptive Deep Networks (ODN) with the purpose of solving the occlusion problem for facial landmark detection. In this model, the occlusion probability of each position in high-level features are inferred by a distillation module that 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, we propose a geometry-aware module to excavate geometric relationships between different facial components. Depending on the synergistic effect of three modules, the proposed network achieves better performance in comparison to state-of-the-art methods on challenging benchmark datasets.

1. Introduction

For many facial analysis tasks, *e.g.*, face recognition [7], face frontalisation [19], and face 3D modeling [26], facial landmark detection is one of pivotal steps, which aims to locate some predefined key-points on facial components. Unfortunately, this significant task still suffers from many challenges in reality, *e.g.*, occlusion, extreme pose, illumination and so on. The occlusion problem is a main obstacle to locate the facial landmarks accurately. Many existing methods [55, 40, 38, 61, 53, 25, 15] perform well for near frontal and untainted face images, while their performances degrade severely if faces undergo severe occlusions. A crucial core of solving the occlusion problem is how to model occlusion. Nevertheless, modeling occlusion explicitly from facial appearance is very difficult because occlusion is irregular, random, and complex.

Recently, some related work has been proposed to solve this challenge. Robust Cascaded Pose Regression (RCPR) [5] divides face into different blocks and explicitly predicts the occlusion likelihood of the corresponding landmarks using a fixed occlusion prior knowledge. However, the training of the RCPR model depends on annotated occlusion state of all the landmarks in the training set. It is very time-consuming to annotate occlusion state of each landmark for large-scale datasets, *e.g.*, 300W [39], AFLW [34], etc. Wu *et al*. [50] leveraged a supervised regression method that gradually updates the landmark visibility by utilizing the appearance, current shape information, and the occlusion consistency. To locate facial landmarks under occlusions, Xing *et al*. [52] introduced an occlusion dictionary into the face appearance dictionary to recover face shape from partially occluded face appearance and model various partial face occlusions. Moreover, Liu *et al.* [32] utilized the shape-indexed appearance to estimate the occlusion level of each landmark, which acts as adaptive weight on the shape-indexed features to decrease the noise on the shape-indexed features.

In recent years, convolutional neural networks (CNN) have been achieved significant performance improvements for facial landmark detection [59, 33, 12, 13, 14, 49]. It is due to the fact that feature extraction process and regression process are trained simultaneously using end-to-end way in CNN that can directly infer the underlying relationship between facial appearance and facial shape. However, occlusion sensitivity is a challenging problem for CNN as well [56]. The occlusion probably mislead CNN on feature representation learning. The localization accuracy would drop significantly if faces are partially occluded.

In this work, we present occlusion-adaptive deep networks (ODN) to overcome the occlusion problem for robust facial landmark detection, which consists of three modules: geometry-aware module, distillation module, and low-rank learning module. First, to model occlusion, the distillation module is used to infer the occlusion probability map based

*Corresponding author.

on high-level features, which serves as the adaptive weight map on high-level features to reduce the impact of occlusion and obtain clean feature representation. Obviously, the clean feature representation cannot represent the holistic face due to the missing semantic features. To obtain exhaustive and complete face feature representation, low-rank learning module is proposed to recover the missing features via learning a *shared structural matrix*. To assist the low-rank learning module to recover lost features, we leverage the geometry-aware module to excavate facial geometric characteristics (*e.g*., symmetry, proximity, position relation, etc.) so that the low-rank module can take advantage of geometric information to better recover lost features. Relying on the synergistic effect of three modules, our proposed ODN can effectively deal with the occlusion problem.

The main contributions in this work are summarized as follows: (1) We present new coherent occlusion-adaptive deep networks to deal with the occlusion problem for facial landmark detection; (2) We suggest a distiller to model occlusion on high-level features implicitly and obtain clean face feature representation; (3) We take advantage a novel module to capture facial geometric characteristics; (4) Low-rank learning is embedded into CNN to recover the missing features and eliminate the redundant features; (5) Experimental results on three challenging benchmark datasets show that our proposed ODN obtains better performance than existing methods.

2. Related Work

In general, existing methods can be categorized into three groups: template methods, coordinate regression methods, and heatmap regression methods.

Template methods. Template models learn a parametric shape model from labeled datasets and exploit Principal Component Analysis (PCA) to model the variation in face shape. Representative work includes Active Contour Model (known as Snakes) [24], Active Shape Model (ASM) [9], Active Appearance Model (AAM) [8], the Constrained Local Model [10], and the Gauss-Newton deformable part models [45]. For this category of algorithms, however, the reconstruction error spreads over the whole face under occlusion [57]. This results in that models cannot accurately locate the landmarks of faces in complex circumstances.

Coordinate regression methods. This category of methods directly learns the mapping from the face images to the landmarks coordinates vectors. Most early work [53, 40, 55, 5, 61] employs the handcrafted features to extract the facial texture information and utilizes SVM, MLP, random forest/fern and so on, as the regressors. For example, SIFT descriptor is used to extract local features of each landmark in SDM [53]. Ren *et al*. [38] proposed local binary feature to capture the local variation of facial appearance. These algorithms usually cascade multi-stages to estimate and update the shape iteratively until convergence. However, the way of prediction in these early work is indirect and sub-optimal because the process of feature extraction and the process of the regression are independent. On the contrary, in recent methods [44, 51, 28, 12, 13, 35, 59], the process of feature extraction and the process of regression are learnt simultaneously in an end-to-end manner. MDM [44] leverages end-to-end recurrent convolutional networks to predict the facial landmarks in coarse-to-fine way. Zhang *et al*. [59] adopted multi-task learning way to regress the coordinates of the facial landmarks and predict the auxiliary attributes simultaneously.

Heatmap regression methods. Heatmap regression methods can be subdivided into two types. The first type usually introduces the landmark heatmaps information to facilitate and guide the learning of network. In Deep Alignment Network (DAN) [27], landmark heatmaps and face images act as the input of intermediate stage in cascaded architecture together and the former can provide visual information about landmark locations. In Look at Boundary (LAB) [48], Wu *et al*. first estimated facial boundary heatmaps and used them to help regress landmarks. Another type of heatmap regression methods directly takes the landmark heapmaps as the groundtruth. Bulat *et al*. [3] proposed a two-stage convolutional part heatmap regression model to settle 3D facial landmark detection. Later, to improve the quality of low resolution facial images and accurately locate the facial landmarks on such poor resolution images, they put forward Super-FAN [4] model that addresses face super-resolution and alignment simultaneously by integrating a heatmap regression based sub-network for facial landmark localization into a super-resolution network.

3. Occlusion-adaptive Deep Networks

In this paper, we propose Occlusion-adaptive Deep Networks (ODN) for facial landmark detection. To be specific, we modify the last residual unit of ResNet-18 [20] to the proposed occlusion-adaptive framework, which aims to effectively cope with the occlusion problem. As illustrated in Fig. 1, occlusion-adaptive framework mainly consists of three close-knit modules: geometry-aware module, distillation module, and low-rank learning module. First, the feature maps $\mathcal{Z}$ from previous residual learning blocks are fed into the geometry-aware module and the distillation module to capture the geometric information and obtain clean feature representation, respectively. Then, the outputs of these two modules are assembled as the input of the low-rank learning module that can recover missing features by modeling the inter-features correlations of faces. We describe each module and the structural relationship among three modules in detail below.

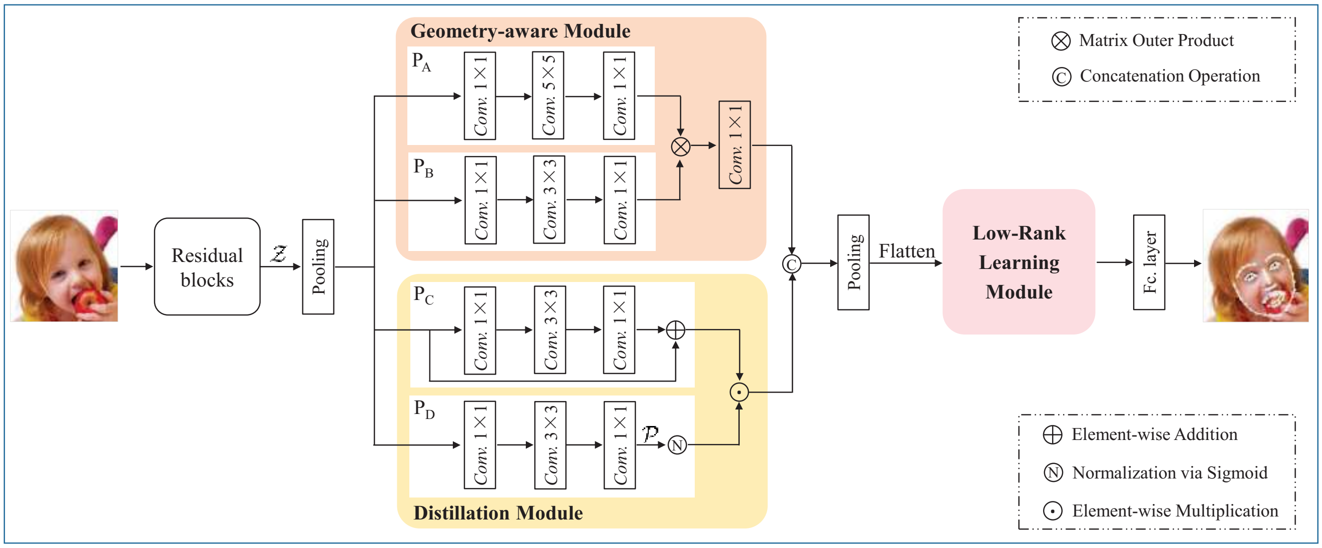
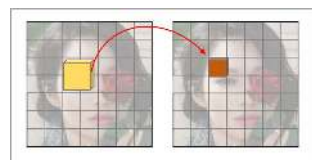


Figure 1. The architecture of Occlusion-adaptive Deep Networks (ODN). Our proposed ODN mainly consists of three modules: geometry-aware module, distillation module, and low-rank learning module.

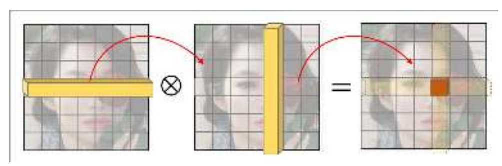
3.1. Geometry-aware Module

As we all know, convolutional operation can only model the relationship in the local neighborhood [47], as shown in Fig. 2(a). Although long-range dependencies can be captured via applying the operation repeatedly, this is computationally inefficient. However, for the task of face landmark detection, geometric relations among different facial components, belonging to long-range dependencies, are also effective information to locate landmarks. Recently, Lin *et al*. [31] proposed to utilize the outer product of the outputs from two CNN streams to obtain pairwise correlations between the feature channels. Inspired by their work, in this paper, we propose a geometry-aware module that exploits the matrix outer product to capture facial geometric relationships among different components.

As shown in Fig. 1, our proposed geometry-aware module is composed of two pathway sub-networks: Pathway-A (P_A) and Pathway-B (P_B). Both of these two pathways are equipped with a 1×1 conv. layer at the front and rear, respectively, which aims to increase the nonlinearity of the decision function without affecting the receptive fields of the conv. layers [41]. To obtain multi-scale features, the middle of the Pathway-A employs a 3×3 conv. layer and the Pathway-B uses a 5×5 conv. layer. The output features of the Pathway-A and the Pathway-B have the same dimension to be compatible. To encode the geometric relations among different facial components, the output features of the two pathways are multiplied to form the high-dimensional geometric feature maps via the matrix outer product of the corresponding channel, as shown in Fig. 2(b). Finally, geometric feature maps are input into a 1×1 conv. layer to get the final geometric representation, which supplies available geometric information for low-rank learning module.



(a) Local receptive field



(b) Our geometry-aware mudule

Figure 2. Comparison of local receptive field and our proposed geometry-aware module on capturing facial geometric relations. $\otimes$ denotes matrix outer product.

In general, element-wise addition and element-wise multiplication are the common ways to aggregate the output features of multiple sub-networks. Element-wise addition often occurs in residual networks family [20, 43] while element-wise multiplication is used to estimate polynomial kernel representation of feature maps [46, 6]. Both of these operations ignore the location of the features and are hence orderless. In this work, the outer product of feature maps is similar to a quadratic kernel expansion, which is indeed a non-local operation to model local pairwise feature interactions for capturing long-range dependencies. It can compute the response at a position as a weighted sum of the features at all positions from the corresponding row and column of the input feature maps.

3.2. Distillation Module

Occluders easily disturb the learning of uncontaminated regions of face and result in failure of convergence during training stage of CNN. To alleviate the sensitivity to occlusion, we propose a distillation module to adaptively filter the features of occluded regions via the self-attention mechanism, even the irrelevant information from background.

Similar to the geometric-aware module, the proposed distillation module also consists of two pathways: Pathway-C (P_C) and Pathway-D (P_D), as displayed in Fig. 1. The Pathway-C exploits a residual block to avoid the decay of input signal, which insures a reliable feature representation. The Pathway-D serves as an occlusion-aware structure to adaptively measure the occlusion probability of each location, which adopts the same '1-3-1' architecture as in Pathway-A. The difference between Pathway-A and Pathway-D is the number of convolutional kernels. It is due to that the Pathway-D only needs less channels to be able to recognize the features of occlusion regions automatically without relying on any specific assumptions. The last 1×1 conv. layer outputs a single-channel feature map $\mathcal{P}$ that is normalized by the following Sigmoid activation function in order to generate a probability map. We integrate this probability map into output feature maps of Pathway-C via element-wise multiplication, aiming to assign small weights to occluded regions and background regions. Hence, we ultimately obtain the clean feature representation (weighted feature maps) of holistic face. Importantly, we take advantage of L_1 *regularization* technique on $\mathcal{P}$ to make it sparse during optimization. Denote $\mathcal{A}$ as the output feature maps of Pathway-C, evidently, it consists of the ideal clean feature representation $\widetilde{A}$ and noise $\mathbb{A}$ (includes background information and occluders). The ideal probability map matrix only has 0 and 1 elements, which can separate the ideal clean data $\widetilde{A}$ from original feature maps $\mathcal{A}$. Thereby, the model ends up using only effective spare features and becomes nearly invariant to the faces with occluders.

Finally, benefited from geometric-aware module and distillation module, geometric feature maps and clean feature representation of holistic face are concatenated into one high-dimensional feature map to generate the hybrid feature representation of face appearance. The hybrid feature maps are down-sampled and flattened into a feature vector as the input of the low-rank learning module.

3.3. Low-rank Learning Module

Although the hybrid features can improve performance, the hybrid features are not exhaustive and complete feature representation of holistic face because the distillation module filters the features of occluded regions. The absence of some features for a face does not necessarily indicate that the face does not have that features, which could be incorrectly interpreted by the model. Since a large number of features/attributes from a face are typically related and co-occur, the presence of some features imply the presence of other features that are closely related, which helps to recover missing features. It is worth noting that our proposed geometry-aware module can provide geometric constraints that are also beneficial to recover missing features. On the other hand, some features may be redundancy information and need to be eliminated. Inspired by [22, 42], we utilize low-rank learning to learn a *shared structural matrix* $\mathcal{M}$ that explicitly encodes the inter-features/attributes correlations so that the missing features can be recovered and the redundant features are removed.

Given the training set$\{(\mathcal{I}_i, \hat{\mathcal{S}}_i)\}_{i=1}^{\mathcal{N}}$, a *shared structural matrix* $\mathcal{M}$ can be learnt to explicitly encode the inter-features/attributes correlations via a rank minimization:

$$\min \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \|\hat{\mathcal{S}}_i - \mathcal{S}\|_F^2 + \beta \mathcal{R}ank(\mathcal{M}), \qquad (1)$$

where the ground-truth of a face is represented as $\hat{\mathcal{S}} = \{s_1, s_2, ..., s_L\}$ and the corresponding prediction is $\mathcal{S}$ ($\mathcal{S} = W_{fc}^{\mathrm{T}} \mathcal{M}^{\mathrm{T}} \mathcal{X}$). Here, $\mathcal{X}$ denotes the hybrid feature vector (outputs of geometry-aware module and distillation module), and W_{fc} is the parameters of fully connection layer (regression coefficient matrix). β is the regularization parameter to control the rank of $\mathcal{M}$ (a larger β induces lower rank). In addition, s denotes single points determined by horizontal and vertical coordinates and L is the number of landmarks for a face. By supervised learning, the structure matrix $\mathcal{M}$ can be learnt in data-driven way to recover missing features effectively by means of helpful geometric information between different facial components.

3.4. Structural Relationship among Three Modules

In our proposed occlusion-adaptive framework, there exists very close-knit relationship among three modules, i.e., geometry-aware module, distillation module, and low-rank learning module. The earlier research [18] showed that the visual processing in the human brain is involved with two streams: the ventral stream and the dorsal stream. The former takes charge of discrimination and recognition of objects while the later processes the object's spatial location information. Similar to this mechanism, our proposed ODN is related to two main information: occlusion-awareness and geometric relationships. To be specific, there exists powerful invariable geometric relationships among different facial components, *e.g.*, symmetry, proximity, position relation and so on, which can be captured by the proposed geometry-aware module. On the other hand, occlusion regions and irrelevant information from background can be filtered by the proposed distillation module. Some lost information from one component can be speculated via other components according to the geometric characteristics. Thereby, the geometric features from the geometry-aware module contributes to the low-rank learning module recovering the missing features on the basis of the clean feature representations from the distillation module. In addition, the relation of opposite and complementary between distillation module and low-rank learning module is benefited to the feature learning of face. From the above, the structural relationship among three modules boosts our proposed ODN to deal with the occlusion problem.

4. End-to-End Optimization

In this section, we introduce how to train our proposed ODN in an end-to-end manner. Mathematically, our proposed ODN can be formulated as the following minimization problem:

$$\begin{aligned}\min \frac{1}{\mathcal{N}}\sum_{i=1}^{\mathcal{N}} \|\hat{\mathcal{S}}_i - \mathcal{S}_i\|_F^2 + \beta \mathcal{R}ank(\mathcal{M}) + \gamma\|\mathcal{M}\|_F^2 \\ +\alpha\|\mathcal{W}_c\|_F^2 + \lambda\|\mathcal{W}_{fc}\|_F^2 + \eta\|\mathcal{P}_i\|_F^1,\end{aligned} \tag{2}$$

where $\mathcal{S} = \mathcal{F}_{ODN}(\mathcal{I}; \mathcal{W}_c; \mathcal{W}_{fc}; \mathcal{M})$. $\mathcal{F}_{ODN}(\cdot)$ denotes our proposed ODN where $\mathcal{W}_c$ and $\mathcal{W}_{fc}$ are the parameter sets of the convolution layers and the fully connection layer, respectively. $\mathcal{M}$ is the parameter set of the low-rank module. Frobenius norms control the shrinkage of three parameter sets with the associated parameters $\{\alpha, \gamma, \lambda\}$, respectively. The single-channel feature map $\mathcal{P}$ from the distillation module is imposed by L_1 *regularization term* with parameter η.

To conduct end-to-end training, the gradients of all terms in (2) should to be derived in the objective function. However, it is an NP-hard problem due to the noncontinuous and non-convex nature of the rank function [60]. The nuclear norm $\|\mathcal{M}\|_*$ is commonly utilized to solve the low-rank learning problem, which provides the tightest lower bound among all convex lower bounds of the rank function. Hence, the objective function (2) can be rewritten as:

$$\begin{aligned}\min \frac{1}{\mathcal{N}}\sum_{i=1}^{\mathcal{N}} \|\hat{\mathcal{S}}_i - \mathcal{S}_i\|_F^2 + \beta \|\mathcal{M}\|_* + \gamma\|\mathcal{M}\|_F^2 \\ +\alpha\|\mathcal{W}_c\|_F^2 + \lambda\|\mathcal{W}_{fc}\|_F^2 + \eta\|\mathcal{P}_i\|_F^1.\end{aligned} \tag{3}$$

By the definition of the nuclear norm [36] and the property of circularity of trace, we can obtain

$$\begin{aligned}\|\mathcal{M}\|_* &= tr(\sqrt{\mathcal{M}^{\mathrm{T}}\mathcal{M}}) = tr(\sqrt{(U\Sigma V^{\mathrm{T}})^{\mathrm{T}}(U\Sigma V^{\mathrm{T}})}) \\ &= tr(\sqrt{V\Sigma^2 V^{\mathrm{T}}}) = tr(\sqrt{VV^{\mathrm{T}}\Sigma^2}) \\ &= tr(\sqrt{\Sigma^2}) = tr(|\Sigma|),\end{aligned} \tag{4}$$

where U, Σ, and V are obtained via the singular value decomposition (SVD) [17] of $\mathcal{M}$. Although the absolute value function $|\Sigma|$ is not differentiable on every point in its domain, we can find a subgradient

$$\frac{\partial\|\mathcal{M}\|_*}{\partial\mathcal{M}} = \frac{\partial tr(|\Sigma|)}{\partial\mathcal{M}} = \frac{tr(\partial|\Sigma|)}{\partial\mathcal{M}} = \frac{tr(|\Sigma|\Sigma^{-1}\partial\Sigma)}{\partial\mathcal{M}}. \tag{5}$$

We know $\mathcal{M} = U\Sigma V^{\mathrm{T}}$ and $\partial\mathcal{M} = \partial U\Sigma V^{\mathrm{T}} + U\partial\Sigma V^{\mathrm{T}} + U\Sigma\partial V^{\mathrm{T}}$. Hence, $U\partial\Sigma V^{\mathrm{T}} = \partial\mathcal{M} - \partial U\Sigma V^{\mathrm{T}} - U\Sigma\partial V^{\mathrm{T}}$. We can get the following equation by multiplying U^{T} on the left side and V on right side of (5), respectively:

$$\partial\Sigma = U^{\mathrm{T}}\partial MV - U^{\mathrm{T}}\partial U\Sigma - \Sigma\partial V^{\mathrm{T}}V, \tag{6}$$

where U is a unitary matrix, *i.e.*, $U^{\mathrm{T}}U = \mathrm{I}$. I is a identity matrix. So, $\partial(U^{\mathrm{T}}U) = \partial\mathrm{I} = \partial U^{\mathrm{T}}U + U^{\mathrm{T}}\partial U = 0$. With the help of this equation, we can compute the rank of second term of (6):

$$\begin{aligned}tr(U^{\mathrm{T}}\partial U\Sigma) &= tr((U^{\mathrm{T}}\partial U\Sigma)^{\mathrm{T}}) = tr(\Sigma^{\mathrm{T}}\partial U^{\mathrm{T}}U) \\ &= -tr(\Sigma U^{\mathrm{T}}\partial U) = -tr(U^{\mathrm{T}}\partial U\Sigma),\end{aligned} \tag{7}$$

which indicates that $tr(U^{\mathrm{T}}\partial U\Sigma) = 0$. Similarly, we also have $tr(\Sigma\partial V^{\mathrm{T}}V) = 0$. Therefore, from (6), we can obtain $tr(\partial\Sigma) = tr(U^{\mathrm{T}}\partial\mathcal{M}V)$. Substituting it into (5), we can have

$$\begin{aligned}\frac{\partial\|\mathcal{M}\|_*}{\partial\mathcal{M}} &= \frac{tr(|\Sigma|\Sigma^{-1}\partial\Sigma)}{\partial\mathcal{M}} = \frac{tr(|\Sigma|\Sigma^{-1}U^{\mathrm{T}}\partial\mathcal{M}V)}{\partial\mathcal{M}} \\ &= \frac{tr(V|\Sigma|\Sigma^{-1}U^{\mathrm{T}}\partial\mathcal{M})}{\partial\mathcal{M}} = (V|\Sigma|\Sigma^{-1}U^{\mathrm{T}})^{\mathrm{T}} \\ &= U\Sigma^{-1}|\Sigma|V^{\mathrm{T}},\end{aligned} \tag{8}$$

as a consequence, we obtain the gradient of the rank function in the objective function.

For the gradients of the first, third, fourth and fifth quadratic term in (2) are easy to calculated. In addition, the last L_1 term also is nondifferentiable, but, we can compute its subdifferential by

$$\frac{\partial\|\mathcal{P}\|_F^1}{p_k} = \begin{cases} \{+1\}, & p_k > 0 \\ \{-1\}, & p_k < 0 \\ [+1, -1], & p_k = 0 \end{cases} \tag{9}$$

where p_k is the k-th element in $\mathcal{P}$.

According to the aforementioned gradient computation equations, our proposed ODN is a directed acyclic graph and the parameters can be learnt in end-to-end way by back-propagating the gradients of the regression loss (*e.g.*, L2 loss).

5. Experiments

Datasets. We evaluate our proposed method on three challenging datasets, including 300W [39], COFW [5], and AFLW [34].

Method	Year	Common set	Fullset
DRMF [1]	2013	6.65	9.22
CFAN [58]	2014	5.50	-
CFSS [61]	2015	4.73	5.99
DR [40]	2016	4.51	6.31
DCRFA [29]	2016	4.19	5.02
RDR [51]	2017	5.05	5.80
SCNN [23]	2017	5.43	6.30
TSR [33]	2017	4.36	4.99
Seq-MT [21]	2018	4.20	4.90
PCD-CNN [28]	2018	**3.67**	**4.44**
ODN		**3.56**	**4.17**

Table 1. Comparison of NRMSE($\times 10^{-2}$) results on Common set and Fullset of 300W.

Method	Year	Challenging set
CMD [55]	2013	19.54
CPR-RPP [54]	2015	11.57
DR [40]	2016	13.80
LBF [38]	2016	11.98
RDR [51]	2017	8.95
DVLN [49]	2017	7.62
TSR [33]	2017	7.56
DSRN [35]	2018	9.68
SBR [13]	2018	7.58
SAN [12]	2018	**6.60**
ODN		**6.67**

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

300W: 300W dataset is a well-known competition dataset for facial landmark detection. Each face is densely annotated with 68 landmarks. It is a collection of 3,837 faces from existing datasets: LFPW [2], AFW [37], HELEN [30], IBUG. We use 3,148 images as training samples and 689 images as testing samples. Specifically, these testing images are split into three subsets: (i) Challenging set (135 images from IBUG); (ii) Common set (554 images, including 224 images from LFPW test set and 330 images from HELEN test set); (iii) Fullset (689 images, containing all of testing images).

COFW: COFW dataset consists of 1,345 images for training and 507 images for test. All training samples are occlusion-free while all testing samples are occluded partially. Each face originally has 29 manually annotated landmarks. For testing set, there is a new version that has been re-annotated with 68 landmarks [16] for purpose of easy comparison to previous methods. In our experiments, we only use testing set with 68 landmarks to verify the effectiveness of dealing with occlusion of our method.

AFLW: AFLW dataset provides a large-scale collection of face images with 21 landmarks for each face, exhibiting a large variety in appearance as well as general imaging and environmental conditions. Following the setting reported in [62], we do not use the landmarks of two ears, and split this dataset into two types: AFLW-Full and AFLW-Frontal. AFLW-Full contains 20,000 training samples and 4,386 testing samples. AFLW-Frontal contains the same training samples as AFLW-Full, but uses 1,165 testing samples with the frontal face.

Evaluation Metric. To evaluate our proposed method, we adopt two evaluation criteria: the normalized root mean squared error (NRMSE), and Cumulative Error Distribution (CED) curve. NRMSE is defined as follow:

$$\mathrm{NRMSE} = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \frac{\| \mathcal{S}_i - \hat{\mathcal{S}}_i \|_2}{L\Omega_i}, \tag{10}$$

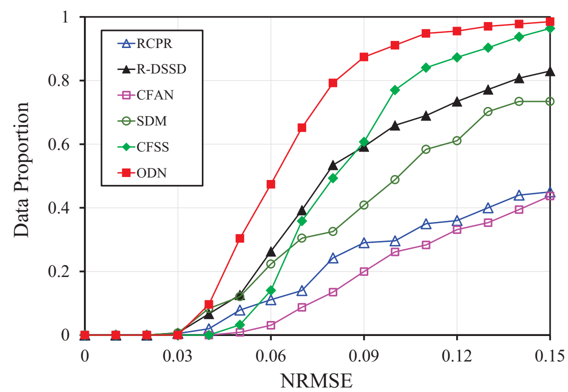


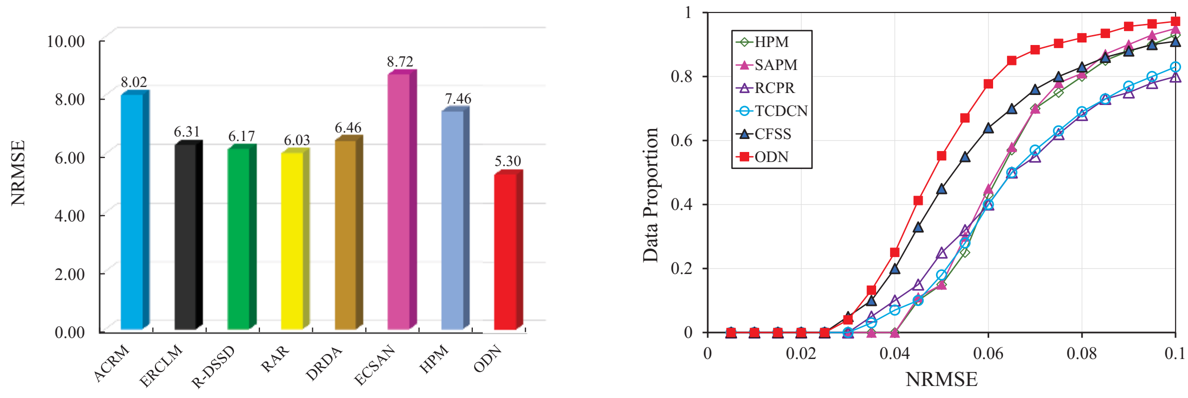
Figure 3. Comparisons of CED curve on Challenging set.

where L, Ω denote the number of landmarks on a face and the inter-ocular distance, respectively. In particular, Ω represents the width of bounding box for AFLW dataset.

Implementation Details. All training images are cropped and resized to 224×224. We exploit rotation, scale, translation and flip operators to conduct data augmentation for training set. In our experiments, all models is pre-trained on the ImageNet dataset [11]. In (2), α, γ, and λ are set to 1×10^{-5}, η and β are set to 1×10^{-6}.

5.1. Evaluation under Normal Circumstances

Firstly, we evaluate the effectiveness of our method on faces under normal circumstances in this subsection. We select two subsets of 300W (Common set and Fullset) as the test datasets. The reason is that most face images in these datasets have less changes under pose, illumination and occlusion. Table 1 shows the experimental results in comparison to the existing benchmarks. From Table 1, we can see that our method outperforms state-of-the-art methods and particularly obtains a good performance gain on Fullset that is already hard to improve. These results indicate that our model can accurately locate landmarks of faces under normal circumstances.



(a) The NRMSE($\times10^{-2}$) criterion

(b) The CED curves criterion

Figure 4. Comparison results of different methods on COFW dataset.

5.2. Evaluation of Robustness against Occlusion

To the best of our knowledge, it is easy for most of state-of-the-art methods to predict the landmarks of normal faces. However, these methods will be in trouble if they attempt to deal with the occlusion problem. Hence, in this subsection, to test the performance of our approach on occluded faces, we conduct the experiments on two difficult datasets: COFW and Challenging set of 300W.

As illustrated in Table 2 and Fig. 3, we compare the proposed method with other representative methods on Challenging set via two kinds of evaluation criteria. The results in Table 2 show that our model boost the NRMSE value to 6.67($\times10^{-2}$), which is competitive with other methods. Note that the NRMSE of DSRN is 9.68($\times10^{-2}$), which also embeds the low-rank learning into CNN. It suggests that our geometry-aware module and distillation module play an important role in boosting the ability to handle the occlusion problem. Furthermore, the Cumulative Error Distribution (CED) curve in Fig. 3 also depicts that our model achieves superior performance in comparison with other methods.

Fig. 4 exhibits the cross-dataset experimental results on COFW dataset that is re-annotated by [16] with 68 landmarks. To be specific, all of models are trained on 300W dataset but evaluated on COFW in order to investigate the robustness of different landmark detection algorithms. As seen in Fig. 4(a), the performance of our proposed ODN greatly exceeds that of other methods. In particular, the NRMSE value of ODN is lower than those methods specific to the occlusion problem, *e.g.*, ACRM, ERCLM, DRDA, and HPM. From the view of another evaluation criterion, Fig. 4(b) demonstrates that the NRMSE value of 92% test samples for our proposed ODN is smaller than 0.08, whereas the highest proportion of other methods is only about 81%. In other words, our method can effectively detect the landmarks for nearly all of test samples from COFW dataset, even if the proposed model is trained on a totally different dataset. Hence, from the experimental results on Challenging set and COFW, we can conclude that our proposed occlusion-adaptive model is robust against occlusion.

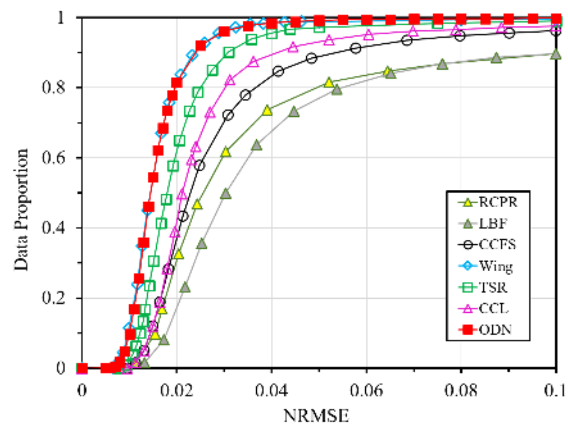


Figure 5. Comparisons of CED curve on AFLW-Full.

5.3. Evaluation of Robustness against Various Poses

Aside from occlusion, extreme pose is also a great challenge for facial landmark detection. To further verify the generalization of our proposed method, we carry out experiments on AFLW dataset that includes a lot of faces with arbitrary pose degree from $-90°$ to $90°$. Two types of performance evaluations for different methods are given in Table 3 and Fig. 5, respectively. In Table 3, our proposed ODN achieves the best score 1.63($\times10^{-2}$) on AFLW-full and 1.38($\times10^{-2}$) on AFLW-Frontal, respectively. We speculate that this mainly owes to our proposed geometry-aware module and low-rank learning module. As we know, many facial geometry characteristics are invariable even if faces undergo arbitrary pose, which can provide the geometric constraints. The geometry-aware module can exactly capture geometric relations among facial components and low-rank learning module is able to employ these relations to recover the lost features. In addition, in Fig. 5, our proposed method almost outperforms others. It is worth men-

Method	SDM [53]	ERT [25]	CCL [62]	DAC-OSR [15]	SBR [13]	SAN [12]	DSRN [35]	ODN
Year	2013	2014	2016	2017	2018	2018	2018	
AFLW-Full	4.05	4.35	2.72	2.27	2.14	1.91	1.86	**1.63**
AFLW-Frontal	2.94	2.75	2.17	1.81	-	1.85	-	**1.38**

Table 3. The NRMSE $(\times 10^{-2})$ comparison of different methods on the AFLW dataset.

Model	NRMSE
BRNet	7.21
BRNet+GM+DM	7.04
BRNet+DM+LM	6.88
BRNet+GM+LM	6.90
BRNet+GM+LM+DM(without L_1)	6.81
BRNet+GM+LM+DM	6.67

Table 4. NRMSE$(\times 10^{-2})$ comparisons of our proposed model with different modules on Challenging set.

tioning that TSR dedicates to solving extreme facial pose problem via using two-stage re-initialization to adjust faces to up-right. But our method dose not adopt any additional measures to adjust facial pose and yet obtains better performance than TSR. These experimental results can prove that our proposed ODN has a great generalization ability to predict the landmarks of faces with arbitrary pose.

5.4. Ablation Study

Our proposed occlusion-adaptive networks consist of three pivotal modules: geometry-aware module (GM), distillation module (DM) and low-rank learning module (LM). In this subsection, we carry out the ablation study to validate their effectiveness on Challenging set. Based on the baseline ResNet-18 (BRNet), we analyse the necessity of existence for each proposed module. Table 4 reports the comparison results of NRMSE.

From Table 4, we can find that each proposed module plays an essential part in improving the performance. Furthermore, it can be obviously observed that the best performance comes from BRNet equipped with three modules simultaneously. Moreover, in our proposed framework, L_1 *regularization* is imposed to single-channel feature map $\mathcal{P}$ and makes it sparse in distillation module. In Table 4, we can see that this regularization operation also obtains small performance gain.

In addition, we show some visualization samples from the distillation module in Fig. 6. The distillation module is related to the self-attention mechanism that can bias the allocation of available processing resources towards the most informative components of an input signal. In Fig. 6, the first column shows face images whose probability maps and post-distilled results are illustrated in the next two columns, respectively. we can see that the distillation module can pay more attention to facial intrinsic regions and reduce the impact of occlusion and background.

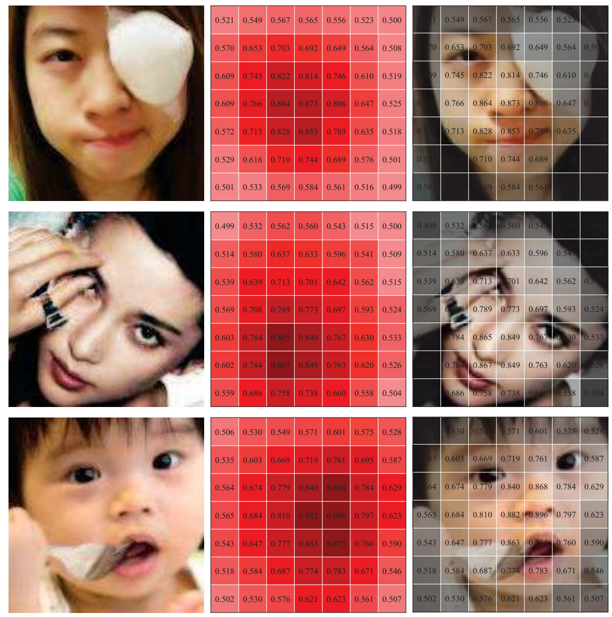


Figure 6. The visualization of some post-distilled face images from COFW dataset

6. Conclusion

In this work, we present an occlusion-adaptive deep network to solve the occlusion problem for facial landmark detection, which is composed of three main modules: geometry-aware module, distillation module and low-rank learning module. Geometry-aware module and distillation module can capture geometric relations of different facial components and obtain clean feature representation, respectively. The outputs of this two modules are concatenated as the input of the low-rank learning module to recover the missing features by means of geometric information.

We conduct 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 pose.

Acknowledgments This work is supported by the National Natural Science Foundation of China Major Program(No. 61827814), the Shenzhen Science and Technology Innovation Commission (SZSTI) project (No. JCYJ20170302153752613) and the National Engineering Laboratory for Big Data System Computing Technology.

References

[1] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Robust discriminative response map fitting with constrained local models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3444–3451, 2013.

[2] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N Kumar. Localizing parts of faces using a consensus of exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 545–552, 2011.

[3] Adrian Bulat and Georgios Tzimiropoulos. Two-stage convolutional part heatmap regression for the 1st 3D face alignment in the wild (3DFAN) challenge. In *European Conference on Computer Vision*, pages 616–624, 2016.

[4] Adrian Bulat and Georgios Tzimiropoulos. Super-FAN: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with GANs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[5] Xavier P. Burgos-Artizzu and Pietro Perona. Robust face landmark estimation under occlusion. In *International Conference on Computer Vision*, pages 1513–1520, 2013.

[6] S. Cai, W. Zuo, and L. Zhang. Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 511–520, 2017.

[7] Dong Chen, Xudong Cao, Fang Wen, and Jian Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3025–3032, 2013.

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

[9] Timothy F. Cootes, Christopher J. Taylor, David H. Cooper, and Jim Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[10] David Cristinacce and Tim Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054–3067, 2008.

[11] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[12] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[13] Xuanyi Dong, Shoou-I Yu, Xinshuo Weng, Shih-En Wei, Yi Yang, and Yaser Sheikh. Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[14] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[15] Zhen-Hua Feng, Josef Kittler, William Christmas, Patrik Huber, and Xiao-Jun Wu. Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3681–3690, 2017.

[16] Golnaz Ghiasi and Charless C. Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1899–1906, 2014.

[17] G. H. Golub and C. Reinsch. Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14(5):403–420, 1970.

[18] Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992.

[19] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. Effective face frontalization in unconstrained images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4295–4304, 2015.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[21] Sina Honari, Pavlo Molchanov, Stephen Tyree, Pascal Vincent, Christopher Pal, and Jan Kautz. Improving landmark localization with semi-supervised learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[22] C. Huang, X. Ding, C. Fang, and D. Wen. Robust image restoration via adaptive low-rank approximation and joint kernel regression. *IEEE Transactions on Image Processing*, 23(12):5284–5297, 2014.

[23] Amin Jourabloo, Mao Ye, Xiaoming Liu, and Liu Ren. Pose-invariant face alignment with a single CNN. In *IEEE International Conference on Computer Vision*, pages 3219–3228. IEEE, 2017.

[24] Michael Kass, Andrew P. Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[25] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

[26] Ira Kemelmacher-Shlizerman and Ronen Basri. 3D face reconstruction from a single image using a single reference face shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):394–405, 2011.

[27] M. Kowalski, J. Naruniec, and T. Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2034–2043, 2017.

[28] Amit Kumar and Rama Chellappa. Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[29] Hanjiang Lai, Shengtao Xiao, Yan Pan, Zhen Cui, Jiashi Feng, Chunyan Xu, Jian Yin, and Shuicheng Yan. Deep recurrent regression for facial landmark detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(5):1144–1157, 2018.

[30] Vuong Le, Jonathan Brandt, Lubomir Bourdev, Lubomir Bourdev, and Thomas S. Huang. Interactive facial feature localization. In *European Conference on Computer Vision*, pages 679–692, 2012.

[31] T. Lin, A. RoyChowdhury, and S. Maji. Bilinear convolutional neural networks for fine-grained visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1309–1322, 2018.

[32] Qingshan Liu, Jiankang Deng, Jing Yang, Guangcan Liu, and Dacheng Tao. Adaptive cascade regression model for robust face alignment. *IEEE Transactions on Image Processing*, 26(2):797–807, 2017.

[33] Jiang-Jing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, Xi Zhou, et al. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3691–3700, 2017.

[34] Peter M. Roth Martin Koestinger, Paul Wohlhart and Horst Bischof. Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. In *IEEE International Conference on Computer Vision Workshops*, pages 2144–2151, 2011.

[35] Xin Miao, Xiantong Zhen, Xianglong Liu, Cheng Deng, Vassilis Athitsos, and Heng Huang. Direct shape regression networks for end-to-end face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[36] Feiping Nie, Heng Huang, and Chris Ding. Low-rank matrix recovery via efficient schatten p-norm minimization. In *the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 655–661, 2012.

[37] Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886, 2012.

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

[39] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2014.

[40] Baoguang Shi, Bai Xiang, Wenyu Liu, and Jingdong Wang. Face alignment with deep regression. *IEEE Transactions on Neural Networks and Learning Systems*, 29(1):183–194, 2018.

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[42] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao. Multi-task learning with low rank attribute embedding for person re-identification. In *IEEE International Conference on Computer Vision*, pages 3739–3747, 2015.

[43] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[44] George Trigeorgis, Patrick Snape, Mihalis A. Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4177–4187, 2016.

[45] Georgios Tzimiropoulos and Maja Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2014.

[46] Hao Wang, Qilong Wang, Mingqi Gao, Peihua Li, and Wangmeng Zuo. Multi-scale location-aware kernel representation for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[47] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803. IEEE, 2018.

[48] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[49] W. Wu and S. Yang. Leveraging intra and inter-dataset variations for robust face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2096–2105, 2017.

[50] Y. Wu and Q. Ji. Robust facial landmark detection under significant head poses and occlusion. In *IEEE International Conference on Computer Vision*, pages 3658–3666, 2015.

[51] Shengtao Xiao, Jiashi Feng, Luoqi Liu, Xuecheng Nie, Wei Wang, Shuicheng Yan, and Ashraf A Kassim. Recurrent 3D-2D dual learning for large-pose facial landmark detection. In *IEEE International Conference on Computer Vision*, pages 1642–1651, 2017.

[52] Junliang Xing, Zhiheng Niu, Junshi Huang, Weiming Hu, Zhou Xi, and Shuicheng Yan. Towards robust and accurate multi-view and partially-occluded face alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):987–1001, 2018.

[53] Xuehan Xiong and Fernando De La Torre. Supervised descent method and its applications to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 532–539, 2013.

[54] H. Yang, X. He, X. Jia, and I. Patras. Robust face alignment under occlusion via regional predictive power estimation. *IEEE Transactions on Image Processing*, 24(8):2393–2403, 2015.

[55] Xiang Yu, Junzhou Huang, Shaoting Zhang, Wang Yan, and Dimitris N Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *IEEE International Conference on Computer Vision*, pages 1944–1951, 2013.

[56] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833, 2014.

[57] Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Occlusion-free face alignment: Deep regression networks coupled with de-corrupt autoencoders. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3428–3437, 2016.

[58] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. *European Conference on Computer Vision*, pages 1–16, 2014.

[59] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5):918–930, 2016.

[60] Xiaowei Zhong, Linli Xu, Yitan Li, Zhiyuan Liu, and Enhong Chen. A nonconvex relaxation approach for rank minimization problems. In *the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 1980–1986, 2015.

[61] Shizhan Zhu, Cheng Li, Change Loy Chen, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4998–5006, 2015.

[62] Shizhan Zhu, Cheng Li, Chen-Change Loy, and Xiaoou Tang. Unconstrained face alignment via cascaded compositional learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3409–3417, 2016.