Region-based Context Enhanced Network for Robust Multiple Face Alignment

Xuxin Lin, Yanyan Liang, *Member, IEEE*, Jun Wan, *Member, IEEE*, Chi Lin and Stan Z. Li, *Fellow, IEEE*

***Abstract*—The recent studies for face alignment have involved developing an isolated algorithm on well-cropped face images. It is difficult to obtain the expected input by using an off-the-shelf face detector in practical applications. In this paper, we attempt to bridge between face detection and face alignment by establishing a novel joint multi-task model, which allows us to simultaneously detect multiple faces and their landmarks on a given scene image. In contrast to the pipeline-based framework by cascading separate models, we aim to propose an end-to-end convolutional network by sharing and transform feature representations between the task-specific modules. To learn a robust landmark estimator for unconstrained face alignment, three types of context enhanced block are designed to encode feature maps with multi-level context, multi-scale context, and global context, respectively. In the post-processing step, we develop a shape reconstruction algorithm based on Point Distribution Model (PDM) to refine the landmark outliers. Extensive experiments demonstrate that our results are robust for the landmark location task and insensitive to the location of estimated face regions. Furthermore, our method significantly outperforms recent state-of-the-art methods on several challenging datasets including 300W, AFLW, and COFW.**

***Index Terms*—facial landmark localization, face alignment, convolutional network, point distribution model.**

I. Introduction

FACE alignment or facial landmark localization, which aims to detect a set of semantic points representing a face shape, has become a widely studied topic in the field of computer vision. It usually serves as a crucial intermediate step in many multimedia and vision applications, such as face distortion recovery [1], emotion recognition [2], [3], and facial expression analysis [4]. In the past two decades, there have been numerous classic methods [5], [6], [7], [8], [9], [10], [11] and annotated common datasets [12], [13], [14], [15], [16], [17] proposed for face alignment. Although an impressive progress has been made for the landmark localization on near-frontal faces, unconstrained face alignment is still a challenging task due to various unexpected conditions in the wild, such as low resolution, large pose, and heavy occlusion.

The previous works can be roughly classified into two categories: template fitting methods [5], [6], [7] and regression-based approaches [8], [9], [10], [11]. The former is to optimize a parametric deformable model to fit a given face image, while the latter extracts features from the image, and regresses a coordinate or response heatmap of each facial landmark directly. To reduce the difference between experimental samples and real-world ones, some in-the-wild face datasets [15], [16], [17] have been proposed to further promote the development of the unconstrained face alignment algorithm. With the increasing number of training samples, the data-driven methods, especially based on Deep Convolutional Neural Network (DCNN) [18], [19], [20], [21] have achieved significant improvements for the landmark localization and led to the state-of-the-art advances. They can learn the discriminative feature representations from a large number of annotated face images by resorting to complex network structures.

Despite most of existing methods perform well on the common face datasets, they usually assume that a well-cropped face image has been acquired by using either the pre-defined bounding box, or the ground-truth landmarks. It might result in that the model easily overfits to the specific face regions and falls down in the real applications, where an off-the-shelf face detector is executed but can not provide the expected bounding boxes. In fact, the problem involves another important vision task: face detection, which estimates the location of face bounding boxes in a given scene image.

In the early years, there have been some studies [12], [22], [23] proposed for jointly face detection and landmark localization. However, their performance is limited by the representation ability of hand-crafted features. Recently, some methods based on DCNN [20], [24], [25], [26] attempted to reduce the performance gap with the existing single-task approaches. Most of them [20], [25], [26] developed a pipeline-based method by cascading multiple independent models. In addition, these methods [24], [25] were proposed to only detect a limited number of facial landmarks. It is difficult to compare them with the existing face alignment algorithms on the popular challenging datasets [15], [16], [17]. Inspired by the recent success of Region-based Convolutional Neural Network (R-CNN) [27], [28], [29] in objection detection, we establish a new R-CNN variant as our detection module. Based on the R-CNN architecture, a new region-wise Fully Convolutional Network (FCN) is designed as the landmark localization module, which can be easily extended in support

Manuscript received October 22, 2018; revised February 06, 2019 and April 25, 2019; accepted May 03, 2019. This work was partially supported by the National Key Research and Development Plan (Grant No. 2016YFC0801002), the Chinese National Natural Science Foundation Projects #61876179, #61872367, Science and Technology Development Fund of Macau (No. 152/2017/A, 0025/2018/A1, 008/2019/A1). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Lei Zhang. *(Corresponding author: Yanyan Liang.)*

X. Lin and Y. Liang are with the Faculty of Information Technology, Macau University of Science and Technology, Macau (e-mail: linxuxin6@gmail.com; yyliang@must.edu.mo).

J. Wan and S. Z. Li are with the Center for Biometrics and Security Research & National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: jun.wan@nlpr.ia.ac.cn; szli@nlpr.ia.ac.cn).

C. Lin is with the University of Southern California, Los Angeles, CA, USA (e-mail: linchi@usc.edu).

of different landmark annotations. In contrast to the cascade structure, we concatenate the feature maps from the task-specific modules to form an end-to-end architecture.

In this paper, we attempt to bridge between face detection and facial landmark localization in an elegant way, where a novel joint multi-task model is proposed to allow us to simultaneously detect multiple faces and their landmarks on a given scene image. The contributions of our work are described as follows:

- We present an end-to-end convolutional network, called Region-based Context Enhanced Network (RCEN), which consists of two task-specific modules. One is a R-CNN variant with two-stage refinement of face bounding boxes, and the other is a region-wise FCN used to learn the response maps of facial landmarks.
- To establish a robust multiple face alignment algorithm, which is less sensitive to various unconstrained conditions, such as low resolution, large pose, and heavy occlusion, we design three types of context enhanced block in RCEN to effectively capture multi-level context, multi-scale context, and global context.
- We develop a PDM-based shape reconstruction algorithm as the post-processing step of RCEN, which exploits the global shape constraint to refine the location of the estimated landmark outliers.
- We demonstrate an advantage of the proposed RCEN over the isolated face alignment algorithm, since it does not depend on the well-cropped face images as input. Furthermore, our approach outperforms recent state-of-the-art results with an obvious margin on several challenging datasets including 300W, AFLW, and COFW.

II. Related Work

In the literature of facial landmark localization, besides some classic template fitting methods (such as ASM [5], AAM [6] and CLM [7]) and traditional regression-based approaches [8], [9], [10], [11], the recent advances have been made by data-driven DCNNs. According to the different purposes, these methods can be categorized as the isolated or multiple face alignment algorithm.

A. Isolated Face Alignment

In the early works, the DCNNs [30], [31] were used to directly learn the mapping from a given face image to the coordinate vector of each landmark. These methods typically cascade multiple convolutional networks to estimate the landmarks in a coarse-to-fine manner. A classic cascaded DCNN was proposed by Sun et al. [30] to design several separate networks responding to different facial parts and regressing the corresponding landmark coordinates. Similarly, Zhang et al. [31] employed multiple auto-encoder networks to process a given facial image with different resolutions and perform the coarse-to-fine face alignment. Recently, Lv et al. [32] presented a deep regression architecture with two-stage re-initialization of a input face image, which is implemented by cascading four sub-networks. Besides the cascaded DCNN, some studies [33], [34] aimed to build an end-to-end DCNN by resorting to well-designed network structures. Zhang et al. [33] formulated a multi-task DCNN that simultaneously learns face alignment and correlated auxiliary tasks for facial attribute analysis. Trigeorgis et al [34] and Hou et al [35] proposed an end-to-end recurrent network to estimate the location of facial landmarks in a coarse-to-fine fashion.

With the appearance of FCN [36], facial landmark localization can be formulated as a heatmap regression problem. Bulat et al. [37] trained a two-stage convolutional aggregation model with deconvolution operation to generate the response map of each landmark. Güler et al. [19] proposed to learn a dense mapping from a face image to a 3D shape template by using a regression-based FCN. Yang et al. [38] employed a stacked hourglass network [39] with a supervised face transformation to predict the landmark heatmaps of normalized faces. Bulat et al. [40] extended the hourglass network using a new bottleneck block to simultaneously consider the 2D and 3D face alignment. In contrast to the coordinate regression, the heatmap regression naturally exploits the spatial structure of feature maps in DCNN, and can be solved better for the fine-grained localization task. Nevertheless, the lack of global context in FCN limits further improvement of the landmark localization capacity. To address that, Zhang et al. [41] and Merget et al. [42] applied the dilated convolution operation into FCN to encode feature maps with large receptive fields. In our network, we specially design a global context enhanced block within a encoder-decoder structure to capture global information.

B. Multiple Face Alignment

In practice, heatmap regression is naturally supported for the landmark localization of multiple faces by setting a confidence threshold to keep multiple landmark sets [42]. However, the relationship among the predicted landmarks in each set is not been established explicitly to represent an individual. Cao et al. [43] proposed a bottom-up landmark localization scheme to model the joint relationship by using part affinity fields. Nevertheless, this work focuses on solving human pose estimation and does not provide more evaluations for facial landmark localization.

In the early works for joint face detection and alignment, Zhu et al. [12] used a mixtures of tree models and HOG features to simultaneously estimate face bounding box, head pose and facial landmarks. Shen et al. [22] proposed an exemplar-based face detector using SIFT features, which integrates a landmark estimator. Chen et al. [23] presented a cascade-based method to jointly handle face detection and alignment by using features of pixel differences. Recently, some methods based on DCNN [24], [25], [26], [44] have aimed to design a unified framework for joint face detection and landmark localization. Zhang et al. [25] cascaded three multi-task DCNNs to estimate face bounding boxes and facial landmarks in a coarse-to-fine manner. Instead of cascading multiple separate models, Chen et al. [24] proposed an end-to-end DCNN by concatenating the feature maps from two task-specific sub-networks. Although these methods are better suited for multiple face alignment, they are limited to the particular scenarios where only a small

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2916455, IEEE Transactions on Multimedia

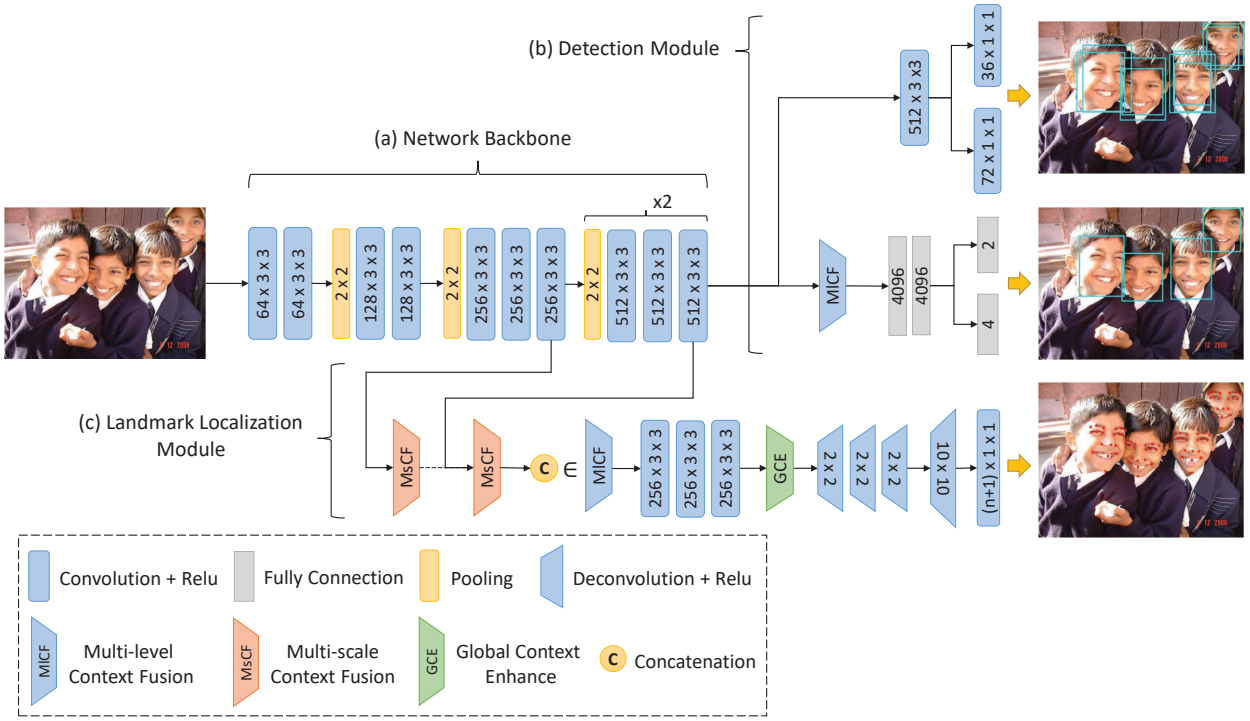


Fig. 1. Overview of the proposed RCEN for multiple face alignment. The network consists of three parts: a) Given a scene image, network backbone generates hierarchical feature maps with different semantic information. b) Detection module receives the image-wise features from network backbone, and predicts the location of each face bounding box. c) Landmark localization module extracts the region-wise features according to the detected face regions, and estimates the corresponding landmark heatmaps. **Note**: the numbers in a "Convolution + Relu" operation denote output number and kernel size, the numbers in "Pooling" and "Deconvolution + Relu" operations denote kernel size, and the number in a "Fully Connection" operation is output number. "c" denotes the concatenate operation in a MlCF block. "n" is the number of predicted facial landmarks.

number of facial landmarks (5 or 21 points) are detected. Hsu et al. [44] used the Faster R-CNN [29] as a face detector and designed a tree structured model for face alignment. Ranjan et al. [26] combined a region proposal algorithm [27] with an extended multi-task DCNN to simultaneously resolve face detection, landmark localization and attribute analysis. By contrast, we aim to build an end-to-end DCNN through sharing and transforming feature representations between the face detection and landmark localization modules.

III. Region-based Context Enhanced Network

As shown in Figure 1, we propose an end-to-end DCNN for joint face detection and landmark localization. The network contains two main modules that are used for the generation of face bounding boxes and the prediction of facial landmark heatmaps, respectively. More detail for the design of our network is discussed in this section. During the training, we employ a two-stage optimization strategy to minimize the multi-task loss function. At the inference, the landmark heatmaps are transformed to corresponding coordinate vectors, and refined by a post-processing step that involves the PDM-based shape reconstruction algorithm.

A. Network Architecture

Overall Structure. In this network, we adopt the VGG-16 [45] as the network backbone containing five convolution stacks, each of which consists of consecutive convolution operations along with non-linear activations. These stacks effectively encode the raw image and generate the hierarchical feature maps with different semantics. From Figure 2, we can see that the feature maps from lower layers tend to respond to the edges and corners of objects, while those from higher layers are class-specific and depict the concerned regions. In the detection module, we follow Faster R-CNN [29] and adopt a typical two-stage detection architecture that consists of a Region Proposal Network (RPN) used to generate candidate targets, and a region-wise convolutional network with fully connected layers for face/non-face decision and bounding box refinement. A crucial difference is that we introduce a multi-level context fusion block into the region-wise network to extract hierarchical features instead of using features only from the top layer of network backbone. For the landmark localization module, we design a new region-wise FCN using encode-decode structure as shown in Figure 1(c). According to the predicted bounding boxes, the module receives the facial features mapped from the hierarchical feature maps, and transforms them to the 120×120 response heatmaps representing the confidence distribution of each landmark as shown in Figure 2.

Context Enhanced Block. In RCEN, we exploit the following three observations to design the context enhanced blocks. 1) In DCNN, the low-layer feature maps are high-resolution, and retain rich texture information with better localization properties, while the high-layer feature maps are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2916455, IEEE Transactions on Multimedia

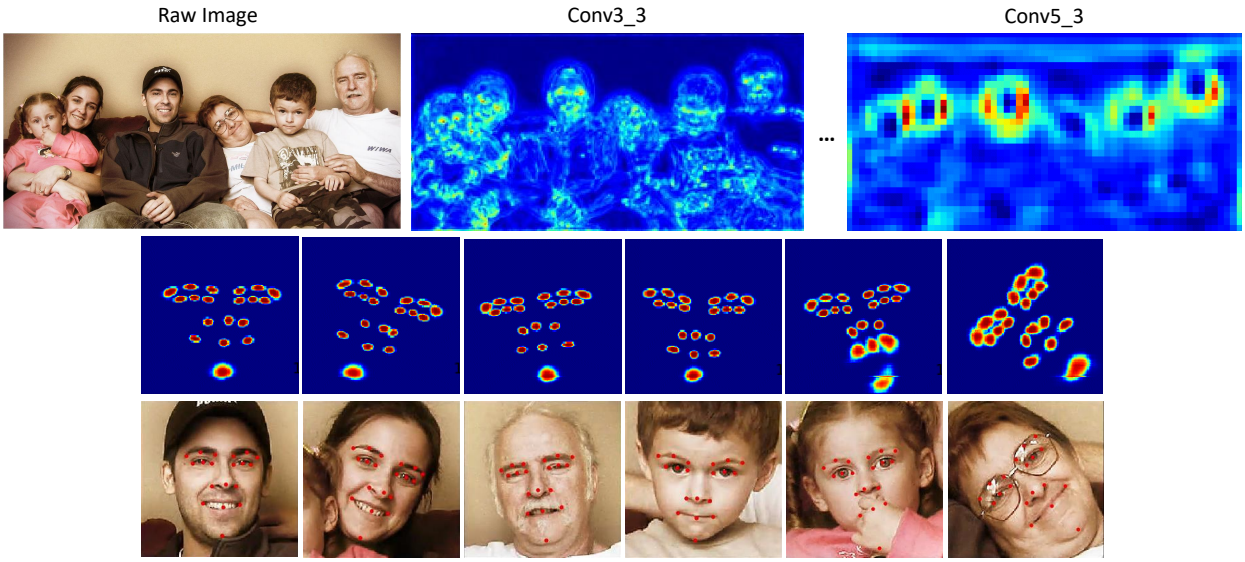


Fig. 2. Top row shows the average activations of feature maps from selected layers (conv3_3 and conv5_3) in network backbone. Middle row shows the sum of all the estimated landmark heatmaps for each detected face. Bottom row shows the corresponding locations of predicted facial landmarks in this image.

semantically strong and suitable for classification tasks [26]. The combination of hierarchical features is natural to provide the multi-level context cues for learning complex tasks like face detection and landmark localization. 2) Due to the limited stride in each convolutional layer, the feature maps only contain single-scale context information, which is not enough to cope with the existence of multi-scale faces in the wild. 3) FCN can not explicitly encode the global context by using the fully connected operation. The lack of global information would affect the inference of FCN especially for the fine-grained tasks like semantic segmentation [46] and landmark localization [42], as the spatial correlation in feature maps can not be exploited effectively with a global view. Recently, there have been some works [47], [48] proposed to combine the multi-level and multi-scale feature maps by using the sparse or dense skip-connections. However, they tend to perform the fusion of global feature maps, which is not necessary since we only consider the local face region. In the following, we will detail three types of context enhanced block that works on a region-wise network to capture multi-level context, multi-scale context and global context, respectively.

- **Multi-level Context Fusion**: Different from the combination of global feature maps from multiple layers, we concatenate the region-wise hierarchical features into a new feature descriptor by using the skip-connections with the Region of Interest (RoI) pooling [29]. Comparing with some existing methods [49], [50] that employ the region-wise feature fusion for improving detection performance, we further develop the operation as a generic block to catch multi-level context in both the detection module and the landmark localization module. As shown in Figure 3(a), according to the position of generated bounding boxes, the feature maps in different layers are first obtained with a fixed size by using RoI pooling. And then, these feature maps are normalized under an unified scale through L2 normalization [51]. Finally, we concatenate all feature maps along the channel axis, and make the cross-channel fusion of them with reduced dimensions by using a convolution operation. In the detection module, the block is grafted after the last three convolution stacks to generate the 7×7 feature maps with 512 channels. For the landmark localization, instead of directly working on the network backbone, the block is used to concatenate the feature maps from three multi-scale context fusion blocks, and generate the 14×14 feature maps with 256 channels.
- **Multi-scale Context Fusion**: Exploiting multi-scale context in DCNN has received increasing attention for object detection [52] and semantic segmentation [53]. The former applies Spatial Pyramid Pooling (SPP) into the single-layer feature maps to create a one-dimensional representation with multi-scale information, while the latter extends SPP further for the building of three-dimensional feature maps adapted to the pixel-wise classification. Inspired by these works, we formulate a Spatial Pyramid RoI Pooling (SPRP) in the new block to handle region-wise feature maps. To our best knowledge, the region-wise multi-scale context is first considered for the landmark localization. As shown in Figure 3(b), we first transform global feature maps to multiple groups of region-wise feature maps with different sizes by using SPRP. And then, expect for the feature maps with the largest size, the others are upsampled by a learned deconvolution operation with non-linear activation. Note that these feature maps keep the same spatial resolution but different receptive fields for a region object. In the landmark localization module, we apply the block into the hierarchical feature maps output from the last three convolution stacks using SPRP with 3 sizes $\{14, 11, 9\}$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2916455, IEEE Transactions on Multimedia

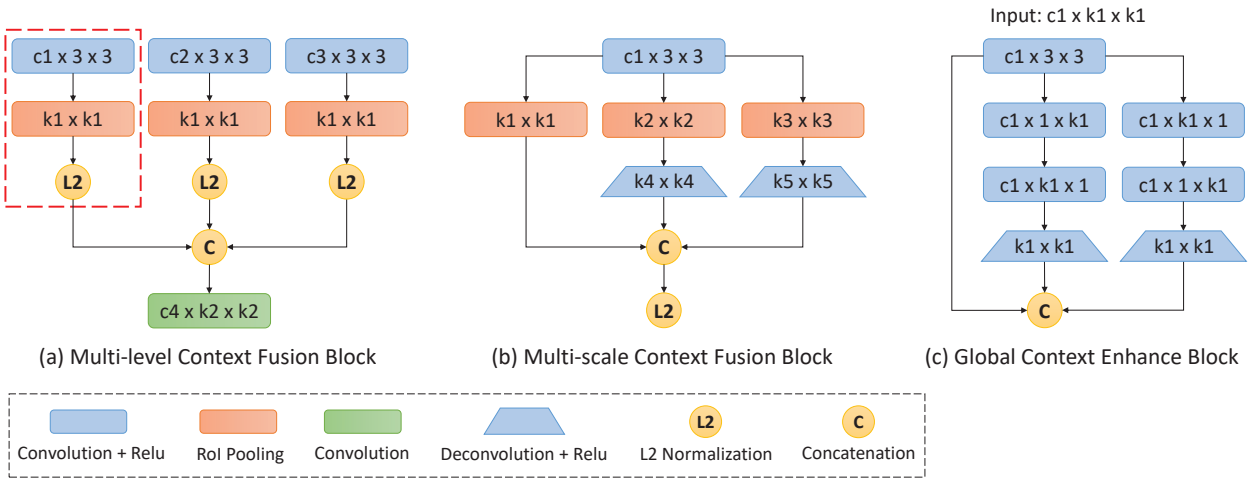


Fig. 3. Three types of context enhanced block to capture multi-level context (a), multi-scale context (b), and global context (c), respectively. **Note**: the red box in multi-level context fusion block represents a single-level context encoding, which is replaced with a multi-scale context fusion block in the landmark localization module. The numbers in a "Convolution + Relu" operation denote output number and kernel size, the numbers in "RoI Pooling" and "Deconvolution + Relu" operations denote pooled size and kernel size, respectively, and the numbers outside denote the channel size and spatial size of input feature maps.

- **Global Context Enhance**: To capture the global context in FCN, there have been some works proposed to employ the stacking dilated convolutions [42] and the separable large filters [46] to encode feature maps with large receptive fields. Nevertheless, the both methods implicitly lose a part of information in either the gaps or the boundaries of the feature maps. In our work, we design a new block based on the residual architecture [54] to encode the region-wise feature maps with a global receptive field. As shown in Figure 3(c), the block consists of a identity mapping and two global feature encoders with the separable convolution operations. As an example, the input feature maps are first encoded along the horizontal and vertical directions by the consecutive $1\times$k1 and k1×1 convolution operations with non-linear activations, respectively, where the k1 is set to the input spatial size for global encoding. And then, the feature maps are upsampled with the original size by a following deconvolution operation. Instead of the element-wise summation used in original residual block [54], we perform the channel-wise concatenation for all the feature maps in the last step. In RCEN, the block is embedded within the encoder-decoder structure from the landmark localization module, and receives the 14×14 feature maps with 256 channels.

B. Training

Bounding Box Normalization. During the training, the positive samples are determined according to their intersection-over-union (IoU) overlaps with the corresponding ground-truth bounding boxes. To unify the distribution of annotated bounding boxes for stable training, we need to normalize the annotations defined on different common datasets. As shown in Figure 4, we can find that the ground-truth bounding boxes are either not provided, or annotated with different styles. In our work, we refer to the definition of annotated faces on the WIDER FACE dataset [55], and

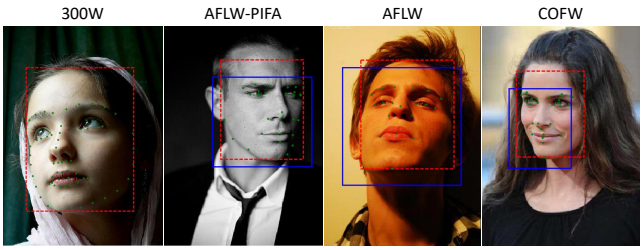


Fig. 4. Different styles of annotation on the 300W, AFLW and COFW datasets. The blue boxes and green points denote the ground-truth bounding boxes and facial landmarks respectively, while the red dotted boxes denote the normalized bounding boxes.

provide new ground-truth bounding boxes using the following steps. 1) An initial bounding box is first obtained with the minimum area according to the coordinates of the outermost landmarks. 2) And then, the height of the bounding box is extended toward the forehead direction with $1/3$ increase. 3) When the contour landmarks are not available on the common datasets such as AFLW-Full [56] and COFW [17], the width of the bounding box is also increased along both sides by $2\times$ distance between the corresponding eyebrow center and outer corner.

Multi-task Loss Function. In RCEN, we consider three functions to minimize: L_{rpn}, L_{det} and L_{lan}, representing the losses of region proposal generation, face detection and facial landmark localization, respectively. Following the work [29], L_{rpn} and L_{det} are defined as follows:

$$\begin{aligned} L_{rpn} &= \sum_i L_{cls}(e_i, e_i^*) + e_i^* \lambda_1 L_{reg}(\mathbf{v}_i, \mathbf{v}_i^*), \\ L_{det} &= \sum_j L_{cls}(h_j, h_j^*) + h_j^* \lambda_2 L_{reg}(\mathbf{w}_j, \mathbf{w}_j^*), \end{aligned} \tag{1}$$

where i is the index of a reference box (anchor) tiled on the output map of RPN. e_i denotes the predicted probability of

the i-th reference box being a face. The ground-truth label e_i^* is 1 if the reference box is positive and 0 otherwise. The vectors $\mathbf{v}_i$ and $\mathbf{v}_i^*$ represent 4 parameterized coordinates of the predicted region proposal and the ground-truth bounding box, respectively. h_j is the predicted probability of the j-th generated region proposal being a face, while h_j^* denotes a ground-truth label that is 1 if the region proposal is positive and 0 otherwise. The vectors $\mathbf{w}_j$ and $\mathbf{w}_j^*$ are the parameterized coordinates of the predicted face and corresponding ground-truth bounding box. The function L_{cls} denotes a softmax loss with respect to two classes (face/non-face) used for classification. The regression loss L_{reg} is the smooth L1 function defined in [28] with a regularization parameter λ.

In our work, the landmark localization is formulated as the pixel-wise classification in $n+1$ response maps using a per-pixel softmax and a multinomial cross-entropy loss, where n is the number of facial landmarks. L_{lan} is defined as follows:

$$L_{lan} = -\sum_{k}\sum_{xy}\sum_{m=0}^{n} I_m(G_k^*(x,y))log\frac{e^{M_k^m(x,y)}}{\sum_{l=0}^{n} e^{M_k^l(x,y)}}, \quad (2)$$

where $G_k^*(x,y) \in \{0,...,n\}$ denotes the ground-truth label of the k-th detected face at pixel location (x,y). I_m is an indicator function such that $I_m(G_k^*(x,y))$ is 1 if $G_k^*(x,y) = m$ and 0 otherwise. It can be used to select the matched response map to calculate the loss according to the ground-truth label. $M_k^m(x,y)$ represents the corresponding score in the m-th response map output from the final convolution layer.

Learning Strategy. Due to the sequence-dependent relationship between the tasks of face detection and landmark localization, it is difficult to jointly optimize the detection module and landmark localization module. To reduce the learning difficulty, we adopt a two-stage optimization strategy to train the proposed RCEN. At the first stage, the weights of network backbone with fully connected layers are first initialized by the pre-trained VGG-16 model [45]. And then, the network backbone and the detection module are fine-tuned by using the training set on WIDER FACE [55], which only provides annotated face bounding boxes. In this process, we assign positive labels to reference boxes and region proposals if their IoU overlaps are higher than 0.5 with the ground-truth bounding boxes. At the second stage, based on the learned weights, we employ a common training set with annotated facial landmarks, described in Section IV-A, to train the landmark localization module with the specific output. Note that the weights of network backbone are not updated in this stage. When the IoU overlaps are higher than 0.7 with the ground-truth bounding boxes, the detected samples are used to calculate the L_{lan} loss. To make the most of the ground-truth landmarks, the bounding box of each positive sample is expanded with the 0.3 increases of the height and width, and transformed to the 120×120 target heatmaps responding to the coordinate of each landmark.

C. PDM-based Shape Reconstruction

Inspired by the work of Point distribution model (PDM) [5], which has been widely used in the classic template fitting approaches [6], [7] to model the shape of an object, we employ it inversely to reconstruct the estimated landmarks by exploiting the global shape correlation. The landmark shape can be denoted as $\mathbf{s} = [\mathbf{s}_1^T, ..., \mathbf{s}_n^T]^T \in \mathbb{R}^{2n\times 1}$ with n landmarks, where $\mathbf{s}_m \in \mathbb{R}^{2\times 1}$ represents the coordinate of m-th landmark. The classic shape model is obtained by Principal Component Analysis (PCA) as follows:

$$\mathbf{p} = \mathbf{U}^T(\mathbf{s} - \bar{\mathbf{s}}), \quad (3)$$

where $\mathbf{p} \in \mathbb{R}^{q\times 1}$ is a shape parameter vector retaining q active components. $\mathbf{U} \in \mathbb{R}^{2n\times q}$ denotes the shape bases with q eigenvectors. $\bar{\mathbf{s}} \in \mathbb{R}^{2n\times 1}$ is the mean shape. In order to remove the similarity transform components from the shapes, the shape model can be augmented with three global transformations (scaling, in-plane rotation and translation) by using Generalized Procrustes Analysis before PCA. This results in a new expression of shape model for each landmark as follows:

$$\mathbf{p}_m = \mathbf{U}_m^T(c\mathbf{R}\mathbf{s}_m + \mathbf{t} - \bar{\mathbf{s}}_m), \quad (4)$$

where c, $\mathbf{R} \in \mathbb{R}^{2\times 2}$ and $\mathbf{t} \in \mathbb{R}^{2\times 1}$ represent the parameters of scale, rotation and translation, respectively, and are used for the global similarity transform. Using the orthonormalization procedure described in [57], the shape model can be briefly denoted as:

$$\mathbf{p}^* = \mathbf{U}^{*T}(\mathbf{s} - \bar{\mathbf{s}}), \quad (5)$$

where $\mathbf{p}^* = (p_1^*, ..., p_4^*, p_1, ..., p_q)^T \in \mathbb{R}^{(4+q)\times 1}$ and $\mathbf{U}^* = (\mathbf{u}_1^*, ..., \mathbf{u}_4^*, \mathbf{u}_1, ..., \mathbf{u}_q) \in \mathbb{R}^{2n\times(4+q)}$ are the concatenation of the similarity parameters p_i^* and similarity bases $\mathbf{u}_i^*$ with above $\mathbf{p}$ and $\mathbf{U}$, respectively. Finally, the reconstructed landmark shape with global rigid transformations is obtained as follows:

$$\hat{\mathbf{s}} = \bar{\mathbf{s}} + \mathbf{U}^*\mathbf{U}^{*T}(\mathbf{s} - \bar{\mathbf{s}}). \quad (6)$$

In our work, we use the reconstructed landmark shape to correct the location of each predicted landmark with the confidence in the response map less than a specific threshold.

Algorithm 1 Proposed Framework for Multiple Face Alignment.

Require:
The input scene image $\mathbf{I}$, the pre-trained RCEN including the network backbone Q, detection module D and landmark localization module L, and the PDM-based shape reconstruction model P.

Output:
The set of face shapes E;

Extracting the basic feature maps $\mathbf{F}$ by $\mathbf{F} = Q(\mathbf{I})$;
Predicting the set of facial bounding boxes B by $B = D(\mathbf{F})$;
for $k = 1$ to K **do**
  Estimating the facial landmark heatmaps $\mathbf{H}_k$ by $\mathbf{H}_k = L(B_k, \mathbf{F})$;
  for $m = 1$ to n **do**
    Obtaining E_k^m by locating the m-th landmark at the peak response position in $\mathbf{H}_k^m$
  end for
  Updating the face shape E_k by $E_k = P(E_k)$;
end for

Fig. 5. Example results of RCEN$_{pd}$, RCEN$_{od}$ and RCEN$_{gt}$ on the 300W fullset, which are shown in top row, middle row and bottom row, respectively.

More detail on the settings of the hyperparameters is discussed in Section IV-C. The pseudo-code in Algorithm 1 shows the inference process of the proposed framework including RCEN and PDM-based shape reconstruction.

IV. Experiment

A. Experimental Setting

Datasets. To demonstrate the effectiveness of our approach for facial landmark localization on various unconstrained faces, we conduct the performance evaluation on three challenging datasets including 300W [15], AFLW [16] and COFW [17].

- **300W** [15]: The 300W dataset has been widely used for the study of facial landmark localization, and provides 68 annotated landmarks for each face. The 300W training set contains $3,148$ images provided in AFW [12], LFPW [13] and HELEN [14]. The 300W test set includes 600 images released in the latest 300W competition [58], and has the same distribution as the IBUG dataset [15]. The 300W fullset contains 689 images for testing, which are split into the common subset with 554 images from LFPW and HELEN, and the challenging subset with 135 images from IBUG.
- **AFLW** [16]: The AFLW dataset contains $24,386$ unconstrained faces with rich pose variations, and provides at most 21 annotated landmarks for each face. From the settings of the works [56], [59], the train-test partition on AFLW is described as follows. The AFLW-Full dataset contains $20,000$ training images and $4,386$ test images with 19 reduced landmarks. The AFLW-Frontal dataset consists of $1,314$ near-frontal images from the AFLW-Full test set. The AFLW-PIFA dataset provides additional 13 landmarks for each face on the subset of AFLW, and contains $3,901$ training images and $1,299$ test images.
- **COFW** [17]: The COFW dataset was proposed for the study of occluded face alignment. It consists of $1,852$ in-the-wild face images with different degrees of occlusion, including $1,345$ training images and 507 test images. In this dataset, each face is annotated with 29 facial landmarks and the corresponding occlusion states.

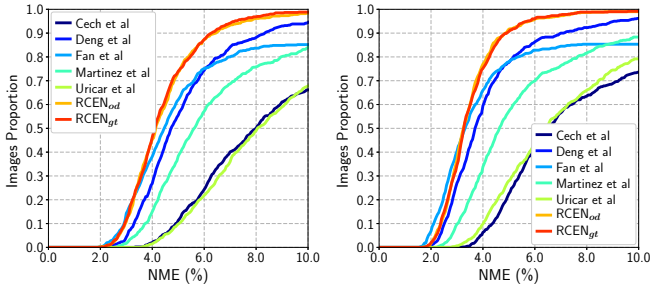


Fig. 6. Comparison of CED curves on the 300W test set with 68 landmarks (a) and 51 landmarks (b).

TABLE I
Comparison of AUC (%) and failure rate (%) on the 300W test set with 68 landmarks and 51 landmarks.

Method	68 points		51 points	
	AUC (%)	Failure rate (%)	AUC (%)	Failure rate (%)
Uricar et al. [60]	21.09	32.17	31.86	20.83
Cech et al. [61]	22.18	33.83	29.51	26.33
Martinez et al. [62]	37.79	16.00	45.80	11.67
Deng et al. [63]	47.52	5.50	57.46	3.83
Fan et al. [64]	48.02	14.83	57.11	14.67
DenseReg [19]	36.05	10.83	-	-
DenseReg+MDM [19]	52.19	3.67	-	-
RCEN$_{od}$	55.29	1.67	**64.75**	1.00
RCEN$_{gt}$	**55.81**	**1.17**	64.55	**0.83**

Evaluation Metric. Following the common evaluation protocols in the previous works, we mainly adopt the point-to-point Normalized Mean Error (NME) in our experiments:

$$NME = \frac{1}{N}\sum_{i=1}^{N}\frac{\|\mathbf{x}_i - \mathbf{x}_i^*\|_2}{d}, \tag{7}$$

where N denotes the number of the tested facial landmarks. d is the normalization factor like inter-ocular distance. $\mathbf{x}_i$ and $\mathbf{x}_i^*$ are the coordinate vectors of the predicted landmark i and the corresponding ground-truth landmark, respectively. Moreover, the Cumulative Error Distribution (CED) curve with the Area-Under-the-Curve (AUC) is provided to describe the relationship between the recall and NME of test samples. The failure rate is used to evaluate the proportion of failure samples with a specific threshold for the maximum NME.

Implementation Details. In our experiments, we follow the train-test partitions on the common datasets, and use the corresponding training set to learn the landmark localization task. Each training sample is first resized so that the shorter side is 600 pixels [28], and then augmented with 3 scale ratios $\{0.5, 1, 2\}$. Note that we do not adopt any spatial transformation for the data augmentation, which has been applied into the existing methods [34], [41], [42]. Based on the setting of the work [29], we use 18 reference boxes with 6 scales $\{16, 32, 64, 128, 256, 512\}$ and 3 aspect ratios

TABLE II
COMPARISON OF NME (%) ON THE 300W COMMON SUBSET, CHALLENGING SUBSET AND FULLSET WITH 68 LANDMARKS.

Method	Common Subset	Challenging Subset	Fullset
CDM [67]	10.10	19.54	11.94
RCPR [17]	6.18	17.26	8.35
CFAN [31]	5.50	16.78	7.69
ESR [8]	5.28	17.00	7.58
SDM [9]	5.57	15.40	7.50
ERT [68]	-	-	6.40
LBF [10]	4.95	11.98	6.32
CFSS [11]	4.73	9.98	5.76
TCDCN [33]	4.80	8.60	5.54
MDM [34]	4.83	10.14	5.88
RAR [69]	4.12	8.35	4.94
FARN [35]	4.23	7.53	4.88
TSR [32]	4.36	7.42	4.96
RDR [70]	5.03	8.95	5.80
CPM [71]	3.39	8.14	4.36
CPM+SBR [71]	3.28	7.58	4.10
SAN [21]	3.34	**6.60**	3.98
RTSM [44]	6.02	16.52	8.06
HyperFace [26]	-	10.88	-
HF-ResNet [26]	-	8.18	-
TSR_{od} [32]	4.36	7.56	4.99
SAN_{od} [21]	3.41	7.55	4.24
RCEN$_{pd}$	3.28	6.73	3.96
RCEN$_{od}$	3.26	6.84	3.96
RCEN$_{gt}$	**3.25**	6.70	**3.93**

TABLE III
COMPARISON OF NME (%) ON THE AFLW-FULL AND AFLW-FRONTAL TEST SETS WITH 19 LANDMARKS.

Method	AFLW-Full	AFLW-Frontal
CDM [67]	5.43	3.77
RCPR [17]	3.73	2.87
SDM [9]	4.05	2.94
ERT [68]	4.35	2.75
LBF [10]	4.25	2.74
CFSS [11]	3.92	2.68
CCL [56]	2.72	2.17
TSR [32]	2.17	-
DAC-CSR [72]	2.27	1.81
CPM [71]	2.33	-
CPM+SBR [71]	2.14	-
SAN [21]	1.91	1.85
RCEN$_{pd}$	2.12	1.70
RCEN$_{od}$	2.11	1.69
RCEN$_{gt}$	**1.78**	**1.62**

$\{0.5, 1, 2\}$ on RPN. During the training, we set a momentum of 0.9 and a weight decay of $5e^{-4}$, and initialize the two-stage learning rates to 0.001 and 0.01, respectively, which are both dropped by 0.1 each 50k iterations until the number of total iterations is 160k. The proposed RCEN is built by using the Caffe framework [65], and the PDM-based shape reconstruction algorithm is implemented based on the Menpo project [66]. The code will be made publicly available.[1]

B. Comparison with Existing Methods

In this subsection, we compare our approach with the existing works, including the typical regression methods like SDM [9], LBF [10] and RCPR [17], and data-driven DCNNs such as CFSS [11], TCDCN [33] and MDM [34]. In addition, several recent state-of-the-art approaches like denseReg [19], CPM [71] and SAN [21], are also reported and discussed in our experiments. For fair comparison, we only use the proposed RCEN at the inference without any post-processing step. Except for the 300W test set, which does not provide bounding boxes for given images, we report the results of RCEN_{pd}, RCEN_{od} and RCEN_{gt} on all the datasets by using provided detected bounding boxes, our detected bounding boxes and ground-truth bounding boxes, respectively.

Evaluation on 300W. We make the evaluation on the 300W test set and fullset. Following the standard setting provided in the latest 300W competition [58], the mean error is normalized by using the inter-ocular distance, and the failure rate is calculated by setting the maximum NME to 0.1. Figure 6 shows the CED curve of each participant in the competition, while the corresponding AUC and failure rate are reported in Table I. Comparing with the previous winners [64], [63], which show the advantages for facial landmark localization with 68 points and 51 points, respectively, our method greatly outperforms all the competitors in both cases. Especially, we achieve the obvious improvements of AUC increased by 3.62%, and failure rate reduced by 2.5% in contrast to the recent state-of-the-art method [19]. The results of RCEN_{od} show the slightly degraded performance, which means that our model is not sensitive to the estimated bounding boxes. For the 300W fullset, we follow the previous works [33], [34], [69], and adopt the same NME metric to evaluate the performance of the proposed RCEN on the common subset, the challenging subset and the fullset. As shown in Table II, RCEN_{pd} shows the promising results comparing with several popular methods using provided bounding boxes, including a related approach [35] using structured RoI pooling and two recent methods [71], [21] that seem to achieve nearly saturated performance. We also report several recent works [44], [26], [32], [21] for joint face detection and alignment, which use a cascaded region proposal method [26] or an external face detector [44], [32], [21] like Faster R-CNN [44]. Comparing to the reduced performance of SAN_{od} [21] using its own detected bounding boxes, RCEN_{od} achieves the similar NMEs on all the sets as RCEN_{pd}. By using the ground-truth bounding boxes, RCEN_{gt} outperforms the both methods with a slight margin. Figure 5 shows the example results of RCEN_{pd}, RCEN_{od} and RCEN_{gt} on the 300W fullset. More example results on 300W are provided in Figure 11.

Evaluation on AFLW. We conduct the experiments on the test sets of AFLW-Full, AFLW-Frontal and AFLW-PIFA to validate the effectiveness of RCEN on face images with various challenging poses. To be consistent with the previous settings [59], [56], [41], the mean error is normalized by the

[1]https://github.com/Linxuxin/RCEN.git

TABLE IV
COMPARISON OF NME (%) ON THE AFLW-PIFA TEST SET WITH 21 LANDMARKS AND 34 LANDMARKS.

Method	21 points (vis.)	34 points (vis.)
CDM [67]	8.59	-
RCPR [17]	7.15	6.26
ERT [68]	7.03	-
SDM [9]	6.96	-
LBF [10]	7.06	-
CFSS [11]	6.75	-
PIFA [59]	6.52	8.04
PAWF [73]	-	4.72
CCL [56]	5.81	-
CALE [37]	2.63	2.96
KEPLER [74]	2.98	-
PIFA-S [75]	-	4.45
DeFA [76]	-	3.86
ECT [41]	3.21	3.36
RCEN$_{pd}$	2.72	2.96
RCEN$_{od}$	2.68	2.96
RCEN$_{gt}$	**2.56**	**2.78**

TABLE V
COMPARISON OF NME (%) AND FAILURE RATE (%) ON THE COFW TEST SET WITH 29 LANDMARKS.

Method	NME (%)	Failure Rate (%)
Human	5.60	-
CDM [67]	13.67	-
ESR [8]	11.20	-
SDM [9]	11.14	-
RCPR [17]	8.50	20.00
OC [80]	7.46	13.24
RPP [81]	7.52	16.20
CCR [82]	7.03	10.90
DRDA [77]	6.46	6.00
Wu et al. [78]	5.93	-
TCDCN [33]	8.05	-
RAR [69]	6.03	4.14
SimLPD [79]	6.40	-
DAC-CSR [72]	6.03	4.73
ECT [41]	5.98	4.54
RCEN$_{pd}$	4.44	1.78
RCEN$_{od}$	4.44	2.56
RCEN$_{gt}$	**4.27**	**1.38**

square root of the bounding box size, and the NME on AFLW-PIFA is calculated only considering the visible landmarks. Table III shows the results of several existing works on AFLW-Full and AFLW-Frontal, including two recent state-of-the-art methods [21], [72], in which the former uses the cropped face images based on ground-truth bounding boxes, and the latter receives the full images with newly estimated bounding boxes. Corresponding to our results, RCEN$_{gt}$ achieves the new state-of-the-art performance with 1.78% and 1.62% NMEs on the both test sets, respectively, while RCEN$_{pd}$ and RCEN$_{od}$ are competitive with the work [21], and better than the method [72]. From Table IV, CALE [37] has outperformed the previous methods [56], [73], [59] by a large margin on the AFLW-PIFA test set. Even for the recently proposed approaches [41], [76], [75], [74], the performance seems to be difficult for further improvement. By contrast, the results of our method are encouraging with the slightly reduced NME by using the ground-truth bounding boxes. Moreover, the performances of RCEN$_{pd}$ and RCEN$_{od}$ are still competitive, and not dramatically degraded due to the deviation of the detected bounding boxes. Figure 12 shows the example results of RCEN$_{od}$ on AFLW.

Evaluation on COFW. To demonstrate the robustness of our method on face images with heavy occlusion, we compare RCEN with the existing methods on the COFW test set. Similar to the setting on 300W, we normalize the mean error by using the inter-ocular distance, and adopt the failure rate with the threshold of 0.1 for the maximum NME. As shown in Table V, several recent methods [77], [78], [79] have been proposed to handle the occluded face alignment, including the state-of-the-art work [78] that almost achieves the human-level performance. In RCEN, we specially design a new block to enhance global context for the inference of occluded facial landmarks. The results of RCEN$_{pd}$ and RCEN$_{od}$ significantly outperform all the reported results. By exploiting the ground-truth bounding boxes, our method achieves further improvement of performance and gets 4.27% NME and 1.38% failure rate. The example results of RCEN$_{od}$ on COFW are shown in Figure 13.

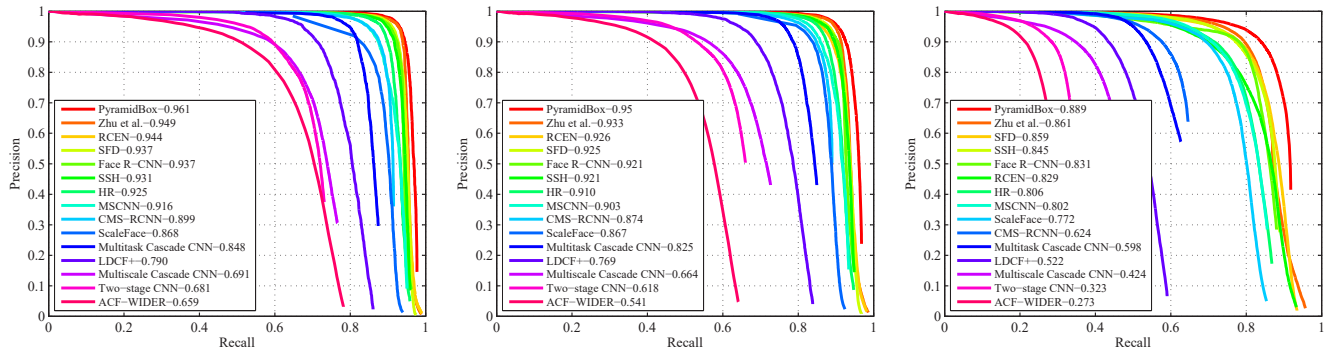
Time Complexity. In the experiments, our method is run on a NVIDIA Tesla K40 GPU and Intel Xeon 3.00GHz CPU. Given a 256×256 face image as input, RCEN$_{od}$ takes about 0.2s with the calculation cost of detection module and landmark localization module. By using the given bounding boxes, RCEN$_{pd}$ and RCEN$_{gt}$ require about 0.08s to run.

C. Component Analysis

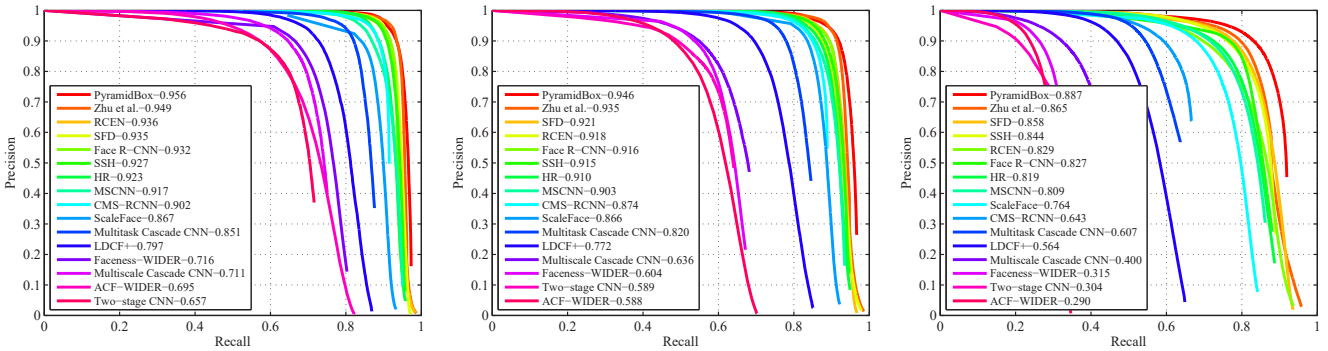
The proposed RCEN consists of several crucial components including the built-in detection module and three types of context enhanced block within the landmark localization module. To validate the effectiveness of these components, we conduct the extended experiments on the WIDER FACE and 300W datasets. In addition, we perform the evaluation of PDM-based shape reconstruction lied on the post-processing step of RCEN by using the 300W and COFW datasets.

Analysis on Detection Module. In contrast to the popular face alignment methods, a crucial difference is that our method exploits a build-in detection module to catch faces, instead of using an off-the-shelf face detector. By following the training setting on WIDER FACE, we compare the proposed RCEN with the recent state-of-the-art face detectors [83], [84], [85], [86], [87], [88], [89] on the WIDER FACE validation and test sets. As shown in Figure 7, our detection module gets competitive performance especially on the easy and medium test sets, and outperforms two related R-CNN variants (CMS-RCNN [83] and Face R-CNN [86]). Our result on the hard set is relatively weak since the detection of tiny face is not our main concern in this paper. In contrast to another work (Multitask Cascade CNN [25]) for joint face detection and alignment, our method achieves significant improvement of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2916455, IEEE Transactions on Multimedia

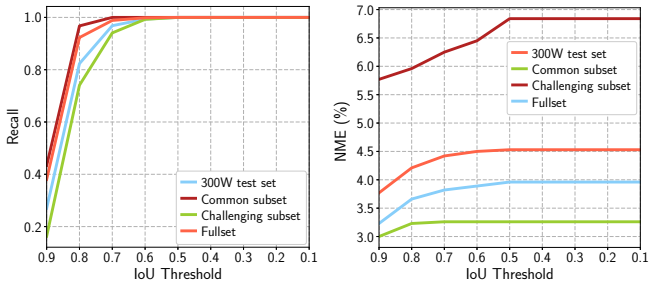


(a) Val: Easy (b) Val: Medium (c) Val: Hard



(d) Test: Easy (e) Test: Medium (f) Test: Hard

Fig. 7. Comparison between the proposed RCEN and other methods on WIDER FACE validation and test sets.



(a) (b)

Fig. 8. Variation of recall (a) and NME (%) (b) with different IoU thresholds on the 300W test set, common subset, challenging subset and fullset.

mAP increased by 8.5%, 9.8% and 22.2% on the easy, medium and hard test sets, respectively.

To further analyze the effect of the detected face regions for the landmark localization performance in RCEN, we evaluate the recall of the detected bounding boxes with the corresponding NME by setting a IoU threshold that determines the minimum IoU overlap of positive samples with an associated ground-truth bounding box. Figure 8 (a) and (b) show the variation of the recall and NME on 300W with different IoU thresholds, respectively. We can find that the recall increases rapidly with the decreasing threshold, and is saturated when the threshold is 0.5. By contrast, the change of the NME is more stable than that of the recall even for the challenging subset where the maximum increase of the NME is about 1%. It means that the performance of the landmark localization module is less sensitive to the location of the estimated bounding boxes, and does not decline sharply even if considering more samples with no well-matched IoU overlaps. For the experiments in Section IV-B, we evaluate all the test samples from each common dataset with a 100% recall by setting a 0.01 IoU threshold.

Analysis on Context Enhanced Block. In the landmark localization module, the context enhanced blocks are designed to capture different context information. To validate the effectiveness of these blocks for improving the landmark localization performance, we compare the proposed RCEN with its three variants by adopting the same setting on the 300W dataset. RCEN-GCE is a reduced RCEN without the use of the global context enhanced block. RCEN-GCE-MsCF is built by removing all the multi-scale context fusion blocks from RCEN-GCE. RCEN-GCE-MsCF-MlCF is a simplest variant of RCEN without using any context enhanced block. As shown in Figure 9, We can find that the performance of RCEN-GCE-MsCF-MlCF is quite poor when the landmark localization module only receives the single feature maps from the last convolution stack. By introducing the multi-level context fusion block, the NME and failure rate of RCEN-GCE-MsCF are both reduced by a large margin. As an example, the both results on the challenging subset are decreased from 12.62% and 51.85% to 7.39% and 14.81%, respectively. By exploiting the multi-scale context and the global context in this network, the performance of RCEN is further improved with the new state-of-the-art results on all the test sets of 300W.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2916455, IEEE Transactions on Multimedia

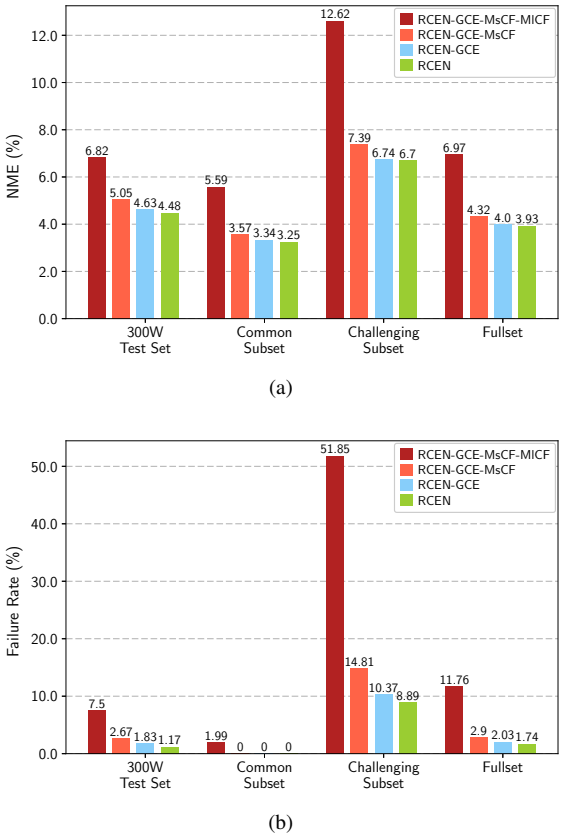


Fig. 9. Comparison of NME (%) (a) and failure rate (%) (b) among RCEN and its three variants on the 300W test set, common subset, challenging subset and fullset.

Analysis on PDM-based Shape Reconstruction. As described in Algorithm 1, the proposed shape reconstruction algorithm is used to refine the estimated landmark outliers from RCEN. The algorithm involves two key hyperparameters, i.e., confidence threshold and number of active components, where the former is used to determine the maximum confidence of the refined landmarks, and the latter denotes the length of the shape parameter vector applied into the shape reconstruction. To estimate the values of them for the inference on a specific common dataset, we first train the PCA model within the algorithm by using the corresponding training set, and then determine the candidate ones by adopting a grid search strategy. As shown in Figure 10, each scattered dot represents a pair of available hyperparameter values with different degrees of NME reduction on the 300W and COFW training sets. Finally, we apply the hyperparameter pair with the minimum NME into the shape reconstruction algorithm on the corresponding test sets. From Figure 10, we can find that the optimal confidence threshold on COFW tends to a higher value than that on 300W, which means that the landmark outliers are more easily induced by the occluded face samples. Table VI reports the performance of RCEN with or without the shape reconstruction algorithm, as well as the corresponding hyperparameter settings on the 300W and COFW test sets. We can see that the NMEs are further reduced after refining the landmark outliers, especially on the 300W challenging subset and the COFW test set.

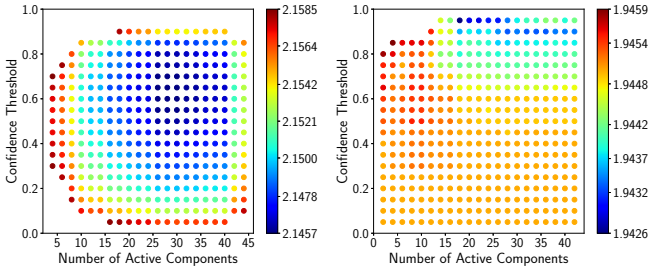


Fig. 10. Variation of NME (%) with different confidence thresholds and numbers of active components on the 300W (a) and COFW (b) training sets. The colors of scattered dots denote the NME values.

TABLE VI
COMPARISON OF NME (%) BETWEEN RCENS WITH OR WITHOUT PDM-BASED SHAPE RECONSTRUCTION (PSR) ON THE 300W TEST SET, COMMON SUBSET, CHALLENGING SUBSET AND FULLSET, AND THE COFW TEST SET.

Test Set	Confidence Threshold	Number of Active Components	w/o PSR (%)	w/ PSR (%)
300W	0.65	30	4.481	4.442
Common Subset	0.65	30	3.251	3.227
Challenge Subset	0.65	30	6.695	6.626
Fullset	0.65	30	3.925	3.893
COFW	0.95	18	4.265	4.174

V. CONCLUSION

In this paper, we propose a novel convolutional network called RCEN for simultaneously detecting multi-face landmarks on a given scene image. Instead of cascading multiple independent models, we aim at establishing an end-to-end joint multi-task model by sharing and transforming feature representations between the task-specific modules. Compared to the isolated face alignment algorithm, a key advantage is that our method does not rely on the well-cropped face images as input, and is insensitive to the estimated bounding boxes from the build-in detection module. With the well-designed context enhanced blocks, the proposed RCEN significantly outperforms recent state-of-the-art methods on several challenging datasets including 300W, AFLW, and COFW. To reduce the performance impact of estimated landmark outliers, a PDM-based shape reconstruction algorithm is presented as the post-processing step of RCEN. In future work, we will extend our model further and validate its effectiveness on other computer vision problems such as multiple human pose estimation and instance segmentation.

REFERENCES

[1] X. Wang, L. Su, H. Qi, Q. Huang, and G. Li, "Face distortion recovery based on online learning database for conversational video," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2130–2140, 2014.

[2] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Two-level hierarchical alignment for semi-coupled hmm-based audiovisual emotion recognition with temporal course," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1880–1895, 2013.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2916455, IEEE Transactions on Multimedia

Fig. 11. Example results of the proposed RCEN on the 300W dataset. Top row shows the images with estimated bounding boxes and facial landmarks on the 300W test set, while middle row and bottom row show the corresponding results on the challenging subset and the common subset, respectively.

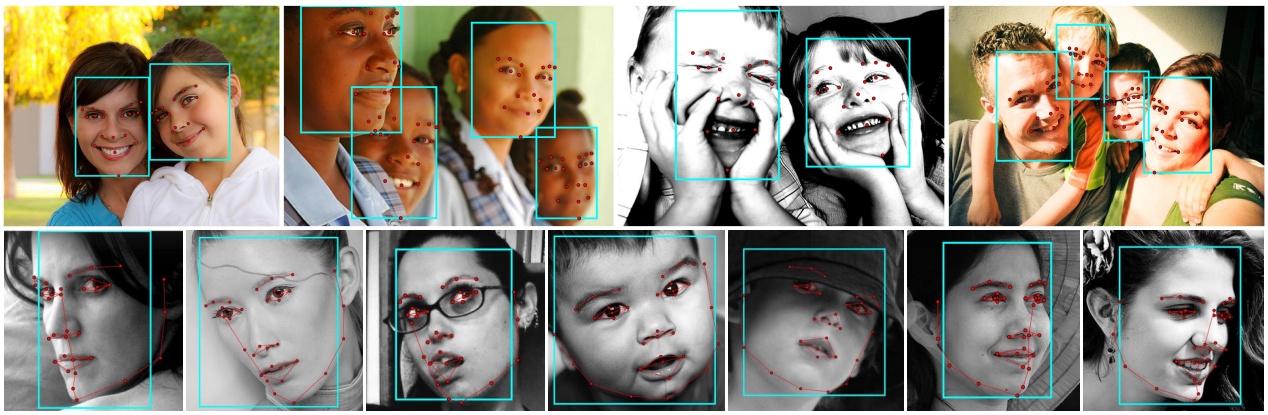


Fig. 12. Example results of the proposed RCEN on the AFLW dataset. The images with estimated bounding boxes and facial landmarks on AFLW-Full and AFLW-PIFA are shown in top row and bottom row, respectively. Note that all the landmarks on AFLW-PIFA are detected from a 3D perspective.

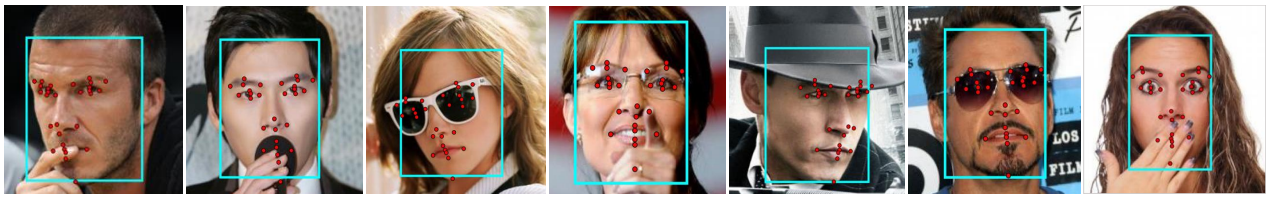


Fig. 13. Example results of the proposed RCEN on the COFW dataset. We show the images with estimated bounding boxes and facial landmarks on the COFW test set.

[3] P. Buitelaar, I. D. Wood, S. Negi, M. Arcan, J. P. McCrae, A. Abele, C. Robin, V. Andryushechkin, H. Ziad, H. Sagha *et al.*, "Mixedemotions: An open-source toolbox for multi-modal emotion analysis," *IEEE Transactions on Multimedia*, 2018.

[4] M. Dahmane and J. Meunier, "Prototype-based modeling for facial expression analysis," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1574–1584, 2014.

[5] T. F. Cootes and C. J. Taylor, "Active shape models-'smart snakes'," in *Proceedings of the British Machine Vision Conference*, 1992, pp. 266–275.

[6] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 681–685, 2001.

[7] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models." in *Proceedings of the British Machine Vision Conference*, 2006, pp. 929–938.

[8] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 177–190, 2014.

[9] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 532–539.

[10] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition, 2014, pp. 1685–1692.

[11] S. Zhu, C. Li, C. Change Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4998–5006.

[12] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2879–2886.

[13] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2930–2940, 2013.

[14] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Proceedings of the European Conference on Computer Vision*, 2012, pp. 679–692.

[15] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 397–403.

[16] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2011, pp. 2144–2151.

[17] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1513–1520.

[18] R. Weng, J. Lu, Y.-P. Tan, and J. Zhou, "Learning cascaded deep auto-encoder networks for face alignment," *IEEE Transactions on Multimedia*, vol. 18, no. 10, pp. 2066–2078, 2016.

[19] R. A. Güler, G. Trigeorgis, E. Antonakos, P. Snape, S. Zafeiriou, and I. Kokkinos, "Densereg: Fully convolutional dense shape regression in-the-wild." in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2017, p. 5.

[20] J. Deng, G. Trigeorgis, Y. Zhou, and S. Zafeiriou, "Joint multi-view face alignment in the wild," *arXiv preprint arXiv:1708.06023*, 2017.

[21] X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Style aggregated network for facial landmark detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2018, p. 6.

[22] X. Shen, Z. Lin, J. Brandt, and Y. Wu, "Detecting and aligning faces by image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3460–3467.

[23] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, "Joint cascade face detection and alignment," in *Proceedings of the European Conference on Computer Vision*, 2014, pp. 109–122.

[24] D. Chen, G. Hua, F. Wen, and J. Sun, "Supervised transformer network for efficient face detection," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 122–138.

[25] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[26] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, 2019.

[27] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[28] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[30] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3476–3483.

[31] J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment," in *Proceedings of the European Conference on Computer Vision*, 2014, pp. 1–16.

[32] J.-J. Lv, X. Shao, J. Xing, C. Cheng, X. Zhou *et al.*, "A deep regression architecture with two-stage re-initialization for high performance facial landmark detection." in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, no. 2, 2017, p. 4.

[33] 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*, vol. 38, no. 5, pp. 918–930, 2016.

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

[35] Q. Hou, J. Wang, R. Bai, S. Zhou, and Y. Gong, "Face alignment recurrent network," *Pattern Recognition*, vol. 74, pp. 448–458, 2018.

[36] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[37] A. Bulat and G. Tzimiropoulos, "Convolutional aggregation of local evidence for large pose face alignment," pp. 86.1–86.12, 2016.

[38] J. Yang, Q. Liu, and K. Zhang, "Stacked hourglass network for robust facial landmark localisation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 2025–2033.

[39] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 483–499.

[40] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 1, no. 2, 2017, p. 4.

[41] H. Zhang, Q. Li, Z. Sun, and Y. Liu, "Combining data-driven and model-driven methods for robust facial landmark detection," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 10, pp. 2409–2422, 2018.

[42] D. Merget, M. Rock, and G. Rigoll, "Robust facial landmark detection via a fully-convolutional local-global context network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 781–790.

[43] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1812.08008*, 2018.

[44] G.-S. Hsu, H.-C. Shie, C.-H. Hsieh, and J.-S. Chan, "Fast landmark localization with 3d component reconstruction and cnn for cross-pose recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 11, pp. 3194–3207, 2018.

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

[46] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matter-simprove semantic segmentation by global convolutional network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1743–1751.

[47] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[48] G. Huang, D. Chen, T. Li, F. Wu, L. Van Der Maaten, and K. Q. Weinberger, "Multi-scale dense convolutional networks for efficient prediction," *arXiv preprint arXiv:1703.09844*, vol. 2, 2017.

[49] S. Bell, C. Lawrence Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2874–2883.

[50] C. Zhu, Y. Zheng, K. Luu, and M. Savvides, "Cms-rcnn: contextual multi-scale region-based cnn for unconstrained face detection," *arXiv preprint arXiv:1606.05413*, 2016.

[51] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proceedings of the European Conference on Computer Vision*, 2014, pp. 346–361.

[53] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[55] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5525–5533.

[56] S. Zhu, C. Li, C.-C. Loy, and X. Tang, "Unconstrained face alignment via cascaded compositional learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3409–3417.

[57] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[58] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image and Vision Computing*, vol. 47, pp. 3–18, 2016.

[59] A. Jourabloo and X. Liu, "Pose-invariant 3d face alignment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3694–3702.

[60] M. Uřičář, V. Franc, D. Thomas, A. Sugimoto, and V. Hlaváč, "Multi-view facial landmark detector learned by the structured output svm," *Image and Vision Computing*, vol. 47, pp. 45–59, 2016.

[61] J. Čech, V. Franc, M. Uřičář, and J. Matas, "Multi-view facial landmark detection by using a 3d shape model," *Image and Vision Computing*, vol. 47, pp. 60–70, 2016.

[62] B. Martinez and M. F. Valstar, "L2, 1-based regression and prediction accumulation across views for robust facial landmark detection," *Image and Vision Computing*, vol. 47, pp. 36–44, 2016.

[63] J. Deng, Q. Liu, J. Yang, and D. Tao, "M3 csr: Multi-view, multi-scale and multi-component cascade shape regression," *Image and Vision Computing*, vol. 47, pp. 19–26, 2016.

[64] H. Fan and E. Zhou, "Approaching human level facial landmark localization by deep learning," *Image and Vision Computing*, vol. 47, pp. 27–35, 2016.

[65] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM International Conference on Multimedia*, 2014, pp. 675–678.

[66] J. Alabort-i Medina, E. Antonakos, J. Booth, P. Snape, and S. Zafeiriou, "Menpo: A comprehensive platform for parametric image alignment and visual deformable models," in *Proceedings of the 22nd ACM International Conference on Multimedia*, 2014, pp. 679–682.

[67] 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 *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1944–1951.

[68] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1867–1874.

[69] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim, "Robust facial landmark detection via recurrent attentive-refinement networks," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 57–72.

[70] S. Xiao, J. Feng, L. Liu, X. Nie, W. Wang, S. Yan, and A. A. Kassim, "Recurrent 3d-2d dual learning for large-pose facial landmark detection." in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1642–1651.

[71] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 360–368.

[72] Z.-H. Feng, J. Kittler, W. Christmas, P. Huber, and X.-J. Wu, "Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2481–2490.

[73] A. Jourabloo and X. Liu, "Large-pose face alignment via cnn-based dense 3d model fitting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4188–4196.

[74] A. Kumar, A. Alavi, and R. Chellappa, "Kepler: keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors," in *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition*, 2017, pp. 258–265.

[75] A. Jourabloo, M. Ye, X. Liu, and L. Ren, "Pose-invariant face alignment with a single cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3219–3228.

[76] Y. Liu, A. Jourabloo, W. Ren, and X. Liu, "Dense face alignment," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 1619–1628.

[77] J. Zhang, M. Kan, S. Shan, and X. Chen, "Occlusion-free face alignment: deep regression networks coupled with de-corrupt autoencoders," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3428–3437.

[78] Y. Wu and Q. Ji, "Robust facial landmark detection under significant head poses and occlusion," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3658–3666.

[79] Y. Wu, C. Gou, and Q. Ji, "Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion," *arXiv preprint arXiv:1709.08130*, 2017.

[80] G. Ghiasi and C. C. Fowlkes, "Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2385–2392.

[81] H. Yang, X. He, X. Jia, and I. Patras, "Robust face alignment under occlusion via regional predictive power estimation," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2393–2403, 2015.

[82] Z.-H. Feng, G. Hu, J. Kittler, W. Christmas, and X.-J. Wu, "Cascaded collaborative regression for robust facial landmark detection trained using a mixture of synthetic and real images with dynamic weighting," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3425–3440, 2015.

[83] C. Zhu, Y. Zheng, K. Luu, and M. Savvides, "Cms-rcnn: contextual multi-scale region-based cnn for unconstrained face detection," in *Deep Learning for Biometrics*. Springer, 2017, pp. 57–79.

[84] P. Hu and D. Ramanan, "Finding tiny faces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1522–1530.

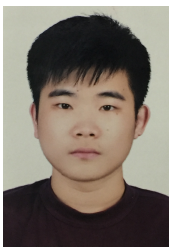
[85] M. Najibi, P. Samangouei, R. Chellappa, and L. Davis, "Ssh: Single stage headless face detector," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4875–4884.

[86] H. Wang, Z. Li, X. Ji, and Y. Wang, "Face r-cnn," *arXiv preprint arXiv:1706.01061*, 2017.

[87] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S^3fd: Single shot scale-invariant face detector," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[88] C. Zhu, R. Tao, K. Luu, and M. Savvides, "Seeing small faces from robust anchors perspective," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5127–5136.

[89] X. Tang, D. K. Du, Z. He, and J. Liu, "Pyramidbox: A context-assisted single shot face detector," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 797–813.



Xuxin Lin received the B.S. degree in software engineering from Beijing Institute of Technology, Zhuhai, China in 2014, and the M.S. degree from Macau University of Science and Technology (MUST) in 2016. Since 2016, he has been a Ph.D. candidate in the MUST. His research interests include computer vision and pattern recognition.

Yanyan Liang received the B.S. degree from Chongqing University of Communication and Posts in 2004, and the M.S. Degree and Ph.D. degree from Macau University of Science and Technology (MUST) in 2006 and 2009. He is currently an assistant professor of MUST. His research is in the fields of computer vision, image processing and Machine Learning.

Jun Wan received the B.S. degree from the China University of Geosciences, Beijing, China, in 2008, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University, Beijing, China, in 2015. Since January 2015, he has been an assistant professor at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA). His main research interests include computer vision, machine learning, especially for gesture and action recognition, facial attribution analysis (i.e. age estimation, facial expression, gender and race classification). He has published papers in top journals, such as JMLR, TPAMI, TIP, TCYB and TOMM. He has served as the reviewer on several top journals and conferences, such as JMLR, TPAMI, TIP, TMM, TSMC, PR, ICPR2016, CVPR2017, ICCV2017, FG2017.

Chi Lin received the B.S. degree with first class honors from the Faculty of Infomation Technology, Macau University of Science and Technology (MUST), Macau, China in 2017. He is currently working toward the M.S. degree at the University of Southern California. From 2015 to 2016, he was selected to participate in "Stars of Tomorrow Internship Program" in Microsoft Research Asia. Since June 2017, he has been a research intern at the Institute of Automation, Chinese Academy of Science (CASIA). His research interests include machine learning, computer vision, and gesture recognition.

Stan Z. Li received the B.Eng. degree from Hunan University, Changsha, China, the M.Eng. degree from National University of Defense Technology, Changsha, China, and the Ph.D. degree from Surrey University, United Kingdom. He is currently a professor and the director of Center for Biometrics and Security Research (CBSR), Institute of Automation, Chinese Academy of Sciences (CASIA). He was at Microsoft Research Asia as a researcher from 2000 to 2004. Prior to that, he was an associate professor at Nanyang Technological University, Singapore. His research interests include pattern recognition and machine learning, image and vision processing, face recognition, biometrics, and intelligent video surveillance. He has published more than 200 papers in international journals and conferences, and authored and edited eight books. He was an associate editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence and is acting as the editor-in-chief for the Encyclopedia of Biometrics. He served as a program cochair for the International Conference on Biometrics 2007 and 2009, and has been involved in organizing other international conferences and workshops in the fields of his research interest. He was elevated to IEEE fellow for his contributions to the fields of face recognition, pattern recognition and computer vision, and he is a member of the IEEE Computer Society.