Facial Landmark Machines: A Backbone-Branches Architecture With Progressive Representation Learning

Lingbo Liu, Guanbin Li, *Member, IEEE*, Yuan Xie, Yizhou Yu, *Fellow, IEEE*, Qing Wang, and Liang Lin, *Senior Member, IEEE*

***Abstract*—Facial landmark localization plays a critical role in face recognition and analysis. In this paper, we propose a novel cascaded backbone-branches fully convolutional neural network (BB-FCN) for rapidly and accurately localizing facial landmarks in unconstrained and cluttered settings. Our proposed BB-FCN generates facial landmark response maps directly from raw images without any preprocessing. BB-FCN follows a coarse-to-fine cascaded pipeline, which consists of a backbone network to roughly detect the locations of all facial landmarks and one branch network for each type of detected landmark to further refine its location. Furthermore, to facilitate the facial landmark localization under unconstrained settings, we propose a large-scale benchmark named SYSU16K, which contains 16 000 faces with large variations in pose, expression, illumination, and resolution. Extensive experimental evaluations demonstrate that our proposed BB-FCN can significantly outperform the state of the art under both constrained (i.e., within detected facial regions only) and unconstrained settings. We further confirm that high-quality facial landmarks localized with our proposed network can also improve the precision and recall of face detection.**

***Index Terms*—Facial landmark localization, cascaded backbone-branches, fully convolutional neural networks, unconstrained settings.**

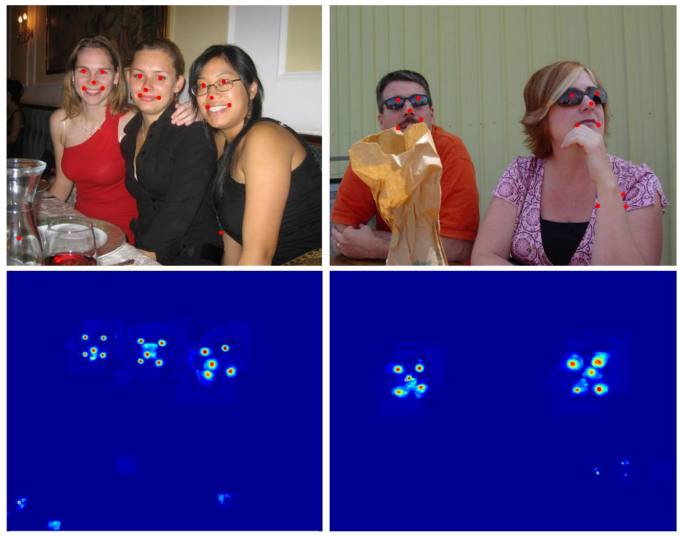


Fig. 1. Facial landmark localization in an unconstrained setting. (a) Two cluttered images with an unknown number of faces. (b) Dense response maps generated by our method.

I. Introduction

FACIAL landmark localization[1] aims to automatically predict key point positions in facial image regions. This task is an essential component in many face-related applications, such as facial attribute analysis [1], face verification [2], [3] and face recognition [4]–[6]. Although tremendous effort has been devoted to this topic, its performance is still far from perfect, particularly on facial regions with severe occlusions or extreme head poses.

Most of the existing approaches for facial landmark localization have been developed for a controlled setting, e.g., the facial regions are detected in a preprocessing step. This setting has drawbacks when we work with images taken in the wild (e.g., cluttered surveillance scenes), where automated face detection is not always reliable. The objective of this work is to propose an effective and efficient facial landmark localization method that is capable of handling images taken in unconstrained settings and that contain multiple faces, extreme head poses and occlusions (see Figure 1). Specifically, we keep the following issues in mind when developing our algorithm.

- Faces may have large appearance and structure variations in unconstrained settings due to diverse viewing conditions, rich facial expressions, large pose changes, facial accessories (e.g., glasses and hats) and aging. Therefore, traditional global models may not work well because the

Manuscript received March 8, 2018; revised September 28, 2018 and November 17, 2018; accepted January 24, 2019. Date of publication February 27, 2019; date of current version August 23, 2019. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFC0830103, in part by the NSFC-Shenzhen Robotics Projects (U1613211), in part by the Hong Kong Research Grants Council under General Research Funds (HKU17206218), in part by the National Natural Science Foundation of China under Grant 61836012, in part by the Ministry of Public Security Science and Technology Police Foundation Project 2016GABJC48, and in part by the SenseTime Research Fund. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Fatih Porikli. *(Corresponding author: Guanbin Li.)*

L. Liu, G. Li, Y. Xie, Q. Wang, and L. Lin are with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China (e-mail: liulingb@mail2.sysu.edu.cn; liguanbin@mail.sysu.edu.cn; xiey39@mail2.sysu.edu.cn; wangq79@mail.sysu.edu.cn; linliang@ieee.org).

Y. Yu is with the Department of Computer Science, The University of Hong Kong, Hong Kong (e-mail: yizhouy@acm.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2019.2902096

[1]Project website: http://www.sysu-hcp.net/facial-landmark-localization/

usual assumptions (e.g., certain spatial layouts) may not hold in such environments.

- Boosted-cascade-based fast face detectors, which evolved from the seminal work of Viola and Jones [7], can only work well for near-frontal faces under normal conditions. Although accurate deformable-part-based models [8] can perform much better on challenging datasets, these models are slow due to their high complexity. Detection in an image takes a few seconds, which makes such detectors impractical for our task.

In this paper, we formulate facial landmark localization as a pixel-labeling problem and develop a fully convolutional neural network (FCN) to overcome the aforementioned issues. The proposed approach produces facial landmark response maps directly from raw images without relying on any preprocessing or feature engineering. Two typical landmark response maps generated with our method are shown in Figure 1.

With the recent advances in deep learning techniques and large-scale annotated image datasets, such as ImageNet, deep convolutional neural network models have achieved significant progress in generic object detection [9], crowd analysis [10], [11] and facial landmark localization [12]. Facial landmark localization is typically formulated as a regression problem. Among the existing methods that take this approach, the cascaded deep convolutional neural networks [13], [14] have emerged as one of the leading methods because of their superior accuracy. Nevertheless, this three-level cascaded CNN framework is complicated and unwieldy. It is arduous to jointly handle the classification (i.e., whether a landmark exists) and localization problems for unconstrained settings. Long *et al.* [15] recently proposed an FCN for pixel labeling, which takes an input image with an arbitrary size and produces a dense label map in the same resolution. This approach shows convincing results for semantic image segmentation and is also very efficient since convolutions are shared among overlapping image patches. Notably, classification and localization can be simultaneously achieved with a dense label map. The success of this work inspires us to adopt an FCN in our task, i.e., pixelwise facial landmark prediction. Nevertheless, a specialized architecture is required because our task requires more accurate prediction than generic image labeling.

Considering both computational efficiency and localization accuracy, we pose facial landmark localization as a cascaded filtering process. In particular, the locations of facial landmarks are first roughly detected in a global context, and then they are refined by observing local regions. To this end, we introduce a novel FCN architecture that naturally follows this coarse-to-fine pipeline. Specifically, our architecture contains one backbone network and several branches, with each branch corresponding to one landmark type. For computational efficiency, the backbone network is designed to be an FCN with lightweight filters, which takes a low-resolution image as its input and rapidly generates an initial multichannel heat map with each channel predicting the location of a specific landmark. We can obtain landmark proposals from each channel of the initial heat map. We then crop a region centered at every landmark proposal from both the original input image and the corresponding channel of the response map. These cropped regions are stacked together and fed to a branch network for a fine and accurate localization. Because fully connected layers are not used in either network, we call our architecture the cascaded backbone-branches fully convolutional network (BB-FCN). Thanks to the tailor-designed architecture of the backbone network, which can reject most background regions and retain high-quality landmark proposals, our BB-FCN is also capable of accurately localizing the landmarks of various scale faces by rapidly scanning every level of the constructed image pyramid. Furthermore, we have also discovered that our landmark localization results can help generate fewer and higher-quality face proposals, thus enhancing the accuracy and efficiency of face detection.

In summary, our contributions in this paper can be summarized as follows:

- We propose a new BB-FCN architecture for facial landmark localization, which consists of a backbone network for rough landmark prediction and a set of branch networks, where each network is for refining the predictions of one specific type of landmark.
- We extensively evaluate BB-FCN on several standard benchmarks (e.g., AFW [8], AFLW [16] and 300W [17]), and our experiments show that BB-FCN achieves superior performance in comparison to other state-of-the-art methods under both constrained (i.e., with face detections) and unconstrained settings. In particular, our BB-FCN significantly decreases the average mean error of the current best-performing method from 8.2% to 6.18% on AFW and from 6.58% to 6.28% on AFLW.
- We use our facial landmark localization results to guide R-CNN-based face detection and demonstrate significant increases in both accuracy and efficiency.

The remainder of this paper is organized as follows. Section II discusses related work and differentiates our method from such works. Section III introduces our proposed BB-FCN architecture. The experimental results and comparisons are presented in Section IV. Finally, Section V concludes this paper.

II. Related Work

Facial landmark localization has long been attempted in computer vision, and a large number of approaches have been proposed for this purpose. The conventional approaches for this task can be divided into two categories: template fitting methods and regression-based methods.

Template fitting methods build face templates to fit input face appearance [18]. A representative work is the active appearance model (AAM) [18], which attempts to estimate model parameters by minimizing the residual between the holistic appearance and an appearance model. A vast collection of methods based on AAM have been proposed [19]–[21]. Rather than using holistic representations, a constrained local model (CLM) [22] learns an independent local detector for each facial keypoint and a shape model for capturing valid facial deformations. Improved versions of CLM primarily differ from each other in terms of local detectors. For instance, Belhumeur *et al.* [23] detected facial landmarks by employing SIFT features and SVM

classifiers, and Liang *et al.* [24] applied AdaBoost to the HAAR wavelet features. These methods are generally superior to the holistic methods due to the robustness of patch detectors against illumination variations and occlusions.

Regression-based facial landmark localization methods can be further divided into direct mapping techniques and cascaded regression models. The former directly maps local or global facial appearances to landmark locations. For example, Dantone *et al.* [25] estimated the absolute coordinates of facial landmarks directly from an ensemble of conditional regression trees trained on facial appearances. Valstar *et al.* [26] applied boosted regression to map the appearances of local image patches to the positions of corresponding facial landmarks. Cascaded regression models [27]–[33] formulate shape estimation as a regression problem and make predictions in a cascaded manner. These models typically start from an initial face shape and iteratively refine the shape according to learned regressors, which map local appearance features to incremental shape adjustments until convergence. Cao *et al.* [27] trained a cascaded nonlinear regression model to infer an entire facial shape from an input image using pairwise pixel-difference features. Burgos-Artizzu *et al.* [34] proposed a novel cascaded regression model for estimating both landmark positions and their occlusions using robust shape-indexed features. Another seminal method is the supervised descent method (SDM) [29], which uses SIFT features extracted around the current shape and minimizes a nonlinear least-squares objective using the learned descent directions. All these methods assume that an initial shape is given in some form, e.g., a mean shape [29], [30]. However, this assumption is too strict and may lead to poor performance on faces with large pose variations.

Despite acknowledged successes, all the aforementioned conventional approaches rely on complicated feature engineering and parameter tuning, which consequently limits their performance in cluttered and diverse settings. Recently, convolutional neural networks and other deep learning models have been successfully applied to various visual computing tasks, including facial landmark estimation. Zhou *et al.* [35] proposed a four-level cascaded regression model based on CNNs, which sequentially predicted landmark coordinates. Zhang *et al.* [12] employed a deep architecture to jointly optimize facial landmark positions with other related tasks, such as pose estimation [36] and facial expression recognition [37]. Zhang *et al.* [38] proposed a new coarse-to-fine DAE pipeline to progressively refine facial landmark locations. In 2016, they further presented de-corrupt autoencoders to automatically recover the genuine appearance of the occluded facial parts, followed by predicting the occlusive facial landmarks [39]. Lai *et al.* [40] proposed an end-to-end CNN architecture to learn highly discriminative shape-indexed features and then refined the shape using the learned deep features via sequential regressions. Merget *et al.* [41] integrated the global context in a fully convolutional network based on dilated convolutions for generating robust features for landmark localization. Bulat *et al.* [42] utilized a facial super-resolution technique to locate the facial landmarks from low-resolution images. Tang *et al.* [43] proposed quantized densely connected U-Nets to largely improve the information flow, which helps to enhance the accuracy of landmark localization. RNN-based models [44]–[46] formulate facial landmark detection as a sequential refinement process in an end-to-end manner. Recently, 3D face models [47]–[51] have also been utilized to accurately locate the landmarks by modeling the structure of facial landmarks. Moreover, many researchers have attempted to adapt some unsupervised [52]–[54] or semisupervised [55] approaches to improve the precision of facial landmark detectors.

Although these methods have achieved remarkable performance, most of them were developed for a controlled setting, which requires a detected frontal face as the input. These methods basically pose landmark estimation as a parameterized regression process, e.g., mapping landmark coordinates, which actually restricts the flexibility in practice due to the fixed form of the parameterization. Such trained models struggle in unconstrained settings (e.g., unknown number of faces in an image). In contrast, our approach produces pixelwise response maps, making it very flexible in localizing facial landmarks in the wild and in integrating with other methods.

III. The Cascaded BB-FCN Architecture

Given an unconstrained image I with an unknown number of faces, our facial landmark localization method aims to locate all facial landmarks in the image. We use $L_i^k = (x_i^k, y_i^k)$ to denote the location of the i^{th} landmark of type k in image I, where x_i^k and y_i^k represent the coordinates of this landmark. Then, our task is to obtain the complete set of landmarks in I,

$$Det(I) = \{(x_i^k, y_i^k)\}_{i,k}, \tag{1}$$

where $k = 1, 2, ..., K$. When describing our method and analyzing the proposed network, we set $K = 5$ as an example, but our method is also applicable to any other values of K. In the experimental section, we will also present simultaneous localization results for 29 landmark types and 68 landmark types. Here, the five landmark types are the left eye (LE), right eye (RE), nose (N), left mouth corner (LM) and right mouth corner (RM).

In contrast to existing approaches that predict landmark locations through coordinate regression, we exploit FCNs to directly produce response maps that indicate the probability of landmark existence at every image location. FCNs have shown excellent performance in various pixel-labeling problems, such as semantic image segmentation [15], object contour detection [56] and salient object detection [57]–[59]. Applying an FCN to an image resembles a deep filtering process. An FCN naturally operates on an input image of any size, and it produces an output with the corresponding spatial dimensions. In our method, the predicted value at each location of the response map can be viewed as a series of filtering operations applied to a specific region of the input image. This specific region is called the receptive field. An ideal series of filters should have the following property: a receptive field with a landmark of a specific type located at its center should return a strong response value, whereas receptive fields without that type of landmark in the center should yield weak responses. Let $F_{\mathbf{W}^k}(P)$ denote the result of applying a series of filtering functions with parameter setting $\mathbf{W}^k$ for type-k

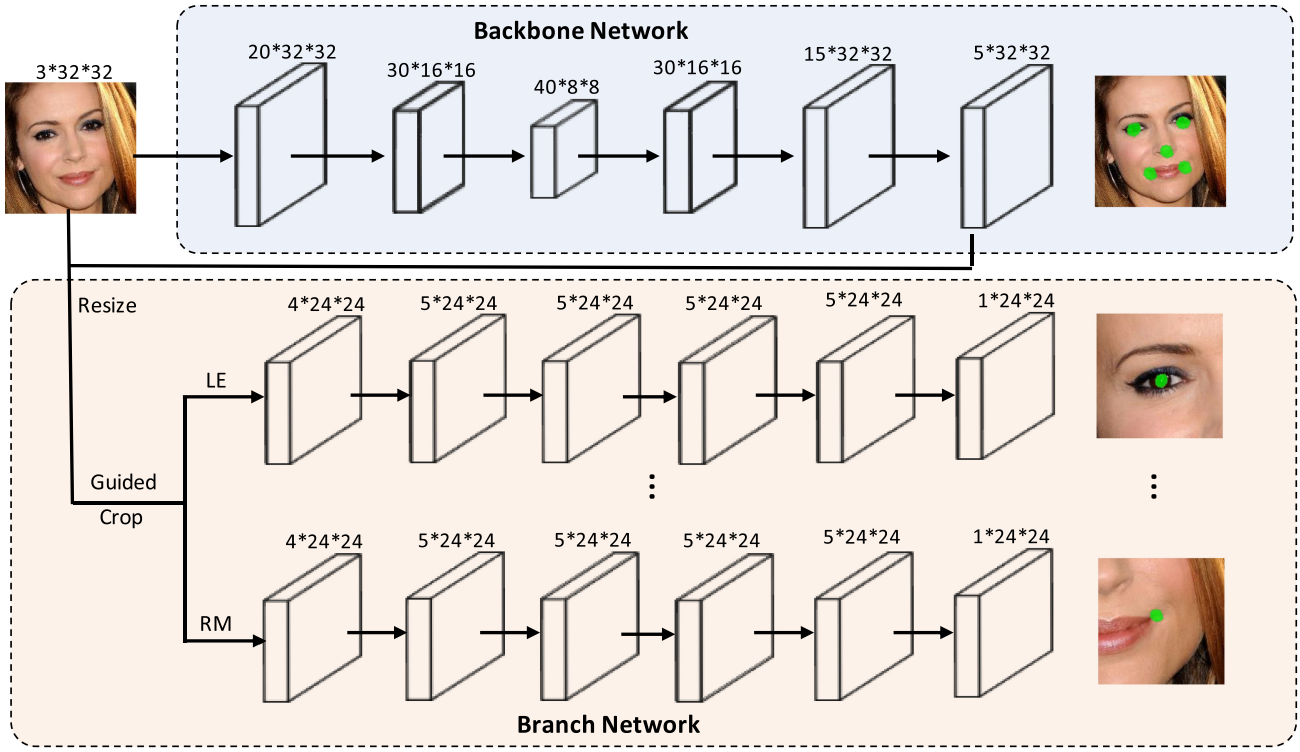


Fig. 2. The main architecture of the proposed backbone-branches fully convolutional neural network. This approach is capable of producing pixelwise facial landmark response maps in a progressive manner. The backbone network first generates low-resolution response maps that identify approximate landmark locations via a fully convolutional network. The branch networks then produce fine response maps over local regions for more accurate landmark localization. There are K (e.g., $K = 5$) branches, each of which corresponds to one type of facial landmark and refines the related response map. Only downsampling, upsampling, and prediction layers are shown, and intermediate convolutional layers are omitted in the network branches.

landmarks to receptive field P, and it is defined as follows:

$$F_{\mathbf{W}^k}(P) = \begin{cases} 1 & \text{if } P \text{ has a type-}k \text{ landmark in the center;} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Applying this function in a sliding window manner to $w \times h$ overlapping receptive fields in an input image I generates a response map $F_{\mathbf{W}^k} * I$ of size $w \times h$, whose value at location (x, y) can be defined as

$$(F_{\mathbf{W}^k} * I)(x, y) = F_{\mathbf{W}^k}(I(P(x, y))), \tag{3}$$

where $I(P(x, y))$ stands for the image patch corresponding to the receptive field of location (x, y) in the output response map.

If the response value is larger than a threshold θ, a landmark of type k is detected at the center of the patch in image I. Thus,

$$Det(I) = \{(\text{center of } P(x, y)) | (F_{\mathbf{W}^k} * I)(x, y) > \theta\}. \tag{4}$$

According to Equation (3), there is a trade-off between localization accuracy and computational cost. To achieve high accuracy, we need to compute response values for significantly overlapping receptive fields. However, to accelerate the detection process, we should generate a coarse response map on less overlapping receptive fields or from a lower-resolution image. This motivates us to develop a cascaded coarse-to-fine process to localize landmarks progressively, in a spirit similar to the hierarchical deep networks in [60] for image classification. Specifically, the architecture of our deep network consists of two components. The first component generates a coarse response map from a relatively low-resolution input, identifying rough landmark locations. Then, the other component takes local patches centered at every estimated landmark location and applies another filtering process to the local patches to obtain a fine response map for accurate landmark localization. This cascaded two-stage strategy enables us to accurately detect facial landmarks at a high speed.

In this paper, this two-component architecture is implemented as a BB-FCN, where the backbone network generates coarse response maps for rough location inference and the branch networks produce fine response maps for accurate location refinement. Figure 2 shows the architecture of our network.

Let a convolutional layer be denoted as $C(n, h \times w \times ch)$ and a deconvolutional layer be denoted as $D(n, h \times w \times ch)$, where n represents the number of kernels and h, w, and ch respectively represent the height, width and number of channels of a kernel. We also use MP to denote a max-pooling layer. In our backbone-branch network, the stride of all convolutional layers is 1, and the stride of all deconvolutional layers is 2. The size of the max-pooling operator is set to 2×2, and the stride for pooling is 2.

A. Backbone Network

The backbone network is an FCN. It can efficiently generate an initial low-resolution response map for input image I. When localizing facial landmarks in an image taken in an

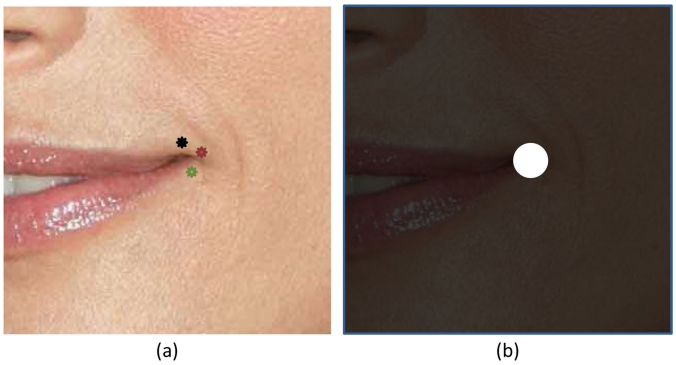


Fig. 3. (a) An isolated point cannot accurately reflect discrepancies among multiple annotations. The three points near the right mouth corner were annotated by three different workers. (b) We label a landmark as a small circular region rather than an as an isolated point in the ground-truth heat map.

unconstrained setting, it can effectively reject a majority of background regions with a threshold. Let $\mathbf{W}_c$ denote its parameters and $H^k(I;\mathbf{W}_c)$ denote the predicted heat map of image I for the k^{th} type of landmarks. The value of $H^k(I;\mathbf{W}_c)$ at position (x,y) can be computed with Equation (3). We train the backbone FCN using the following loss function:

$$\mathcal{L}_1(I;\mathbf{W}_c) = \sum_{k=1}^{K} ||H^k(I;\mathbf{W}_c) - H_c^k(I)||^2, \tag{5}$$

where $H_c^k(I)$ denotes the ground-truth heat map for type-k landmarks.

The backbone network is trained with a patch-based optimization scheme. During the training phase, the human faces are cropped from the unconstrained crowded images and resized to a low resolution of 32×32. Taking the cropped patches of whole faces as input, the backbone network can implicitly learn the geometric constraints among landmarks and generate the response heat maps of all facial landmarks together. Specifically, the backbone network consists of eight convolutional layers with lightweight filters and two deconvolutional layers, which are detailed as follows: $C(20, 5\times5\times3)$ - $C(20, 5\times5\times20)$ - MP - $C(30, 5\times5\times20)$ - $C(30, 5\times5\times30)$ - MP - $C(40, 5\times5\times30)$ - $C(40, 5\times5\times40)$ - $D(30, 2\times2\times40)$ - $C(30, 5\times5\times30)$ - $D(15, 2\times2\times30)$ - $C(5, 1\times1\times15)$.

B. Branch Network

The branch network is composed of K branches, with each branch responsible for detecting one type of landmark. All the K branches are designed to share the same network structure. In branch networks, a cropped patch from the original input image and a region from the backbone's output heat map are stacked together as its input. Therefore, the input data consist of four channels, including 3 channels from the original RGB image and 1 channel from the corresponding channel of the backbone's output heat map. To make the branch network better suited for landmark position refinement, we resize the original input image to 64×64, which is four times the size of the backbone's input, and simultaneously magnify the heat map from the backbone network to 64×64. The resolution of all the cropped patches is 24×24, and they are all centered at the landmark position predicted by the backbone network. As shown in Fig. 2, each branch is trained in the same way as the backbone network. We denote the parameters of the branch component for type-k landmarks as $\mathbf{W}_f^k$, and we respectively use $H(P;\mathbf{W}_f^k), H_0^k(P)$ to denote the predicted fine heat map and the corresponding ground-truth heat map of patch P. The loss function of this branch component is again defined as follows:

$$\mathcal{L}_2(P;\mathbf{W}_f^k) = ||H(P;\mathbf{W}_f^k) - H_0^k(P)||^2. \tag{6}$$

Each branch component is composed of 5 convolutional layers without any pooling operations. The dimensionality of its input data is $24 \times 24 \times 4$. The first 4 convolutional layers consist of 5 channels with the kernel size equal to 5 and stride equal to 1, while the last convolutional layer consists of 5 channels with a kernel size of 1 and stride of 1. As shown in Figure 2, each branch FCN component is detailed as follows: $C(5, 5\times5\times4)$ - $C(5, 5\times5\times5)$ - $C(5, 5\times5\times5)$ - $C(5, 5\times5\times5)$ - $C(1, 1\times1\times5)$.

C. Ground-Truth Heat Map Generation

To our knowledge, the ground truth of a facial landmark is traditionally given as a single pixel location (x,y) in all public datasets. To adapt such landmark specifications for the training stage of our proposed BB-FCN network, we generate the ground-truth heat map of an input image according to the annotated facial landmark locations. The most straightforward method assigns "1" to a single pixel corresponding to each landmark location and "0" to the remaining pixels. However, we argue that this method is suboptimal because an isolated point cannot reflect discrepancies among multiple annotations. As shown in Figure 4(a), the right mouth corner has three slightly different locations marked by three annotators. To take such discrepancies into consideration, we label each landmark as a small region rather than as an isolated point. We first initialize the heat map with zeros everywhere, and then for each landmark p, we mark a circular region with center p and radius R in the ground-truth heat map with 1. Different radii are adopted for the backbone network and branch networks, denoted as R_c and R_f, respectively. R_f is set to be smaller than R_c because the backbone network estimates coarse landmark positions while the branch networks predict accurate landmark locations.

D. Selective Response Map Training

According to Equations (5) and (6), the loss is computed over the full response map. However, this approach gives rise to a severe imbalance between positive and negative training samples because landmarks are very sparse. This unbalanced setting could mislead the response map to take all zero values when the loss is minimized. Therefore, we adopt a selective scheme, i.e., randomly choosing the same number of non-landmark locations as landmark locations in the ground-truth response map to propagate the errors while inhibiting all other non-landmark locations during error backpropagation. For some invisible landmarks or background images, the ground-truth maps have no

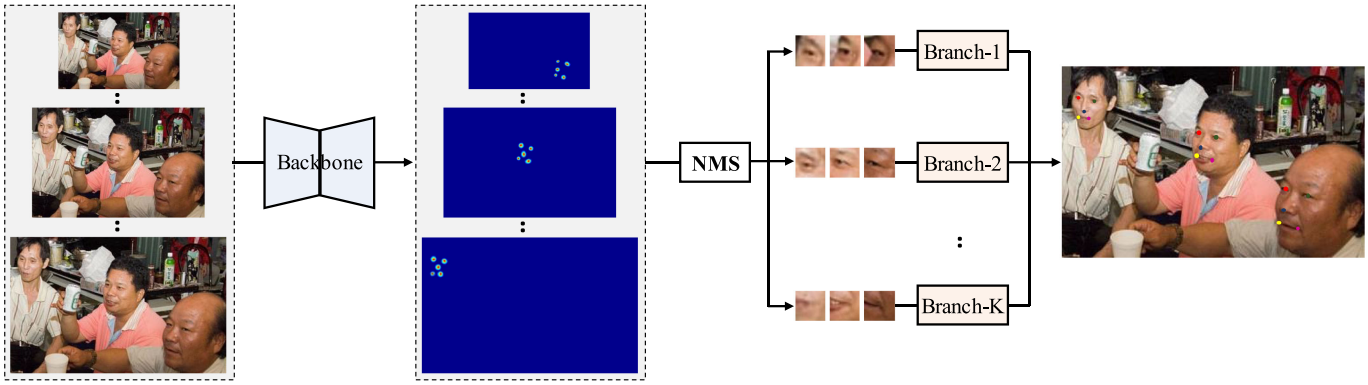


Fig. 4. Illustration of the facial landmark testing procedure under an unconstrained setting. Given an unconstrained image, we first construct an image pyramid. Then, we feed the images at different levels of the pyramid to the backbone network for generating the landmark candidate regions. After adopting a nonmaximum suppression (NMS) to reduce the highly overlapping regions, we refine the locations of the remaining candidate regions with the branch networks. Best viewed in color with magnification.

positive region, and we only select a small ratio of the non-landmark locations to propagate. This selective training scheme is critical in ensuring the convergence of training sessions in our experiments. In addition, for more effective training and more precise results, hard negative mining is also employed. In the selective phase, hard negative samples, which are non-landmark locations with large output values, are selected to propagate the errors when the loss on the validation set stops decreasing.

E. Implementation Details

We have implemented our proposed BB-FCN network in Caffe. A GTX Titan X GPU is used for both training and testing. During training, we randomly initialize our networks by drawing weights from a zero-mean Gaussian distribution with a standard deviation equal to 0.01. The size of a minibatch is set to 40, and the ratio between the numbers of positive and negative training images in each batch is $1:1$ for the backbone network and $4:1$ for the branch networks. The positive training images are image regions cropped from face images in our SYSU16K dataset, which will be described in Section IV-A. The intersection-over-union (IoU) between any cropped region and the original face image is above 0.5. The negative training samples are nonfacial regions randomly cropped from the Pascal VOC 2012 dataset [61]. Both the backbone and branch networks are trained using backpropagation and stochastic gradient descent (SGD) with the momentum set to 0.9 and weight decay set to 0.0005. When training the backbone network, we set the learning rate to 0.001 and the total number of iterations to 25K. The radius R_c of landmark circles is set to 5% of the width of the input image. For the branch networks, the total number of iterations is set to 50K. The learning rate is set to 10^{-4} for the first 30K iterations and 10^{-5} for the last 20K iterations. The radius R_f of landmark circles is set to 3% of the width of the input image. During training, only a subset of the non-landmark locations in the heat map are chosen to propagate errors, as described in Section III-D.

During the testing phase, our BB-FCN network is able to accurately locate facial landmarks under both constrained and unconstrained settings. For convenience in the following part, we denote the average position of the n locations with the highest response values in a 2D heat map M as $Ave\{M, n\}$.

1) Constrained Setting: Given a cropped facial image I, we first resize it to 32×32 and feed it to the backbone network to generate the coarse response heat map M_c. Because the radius R_c is set to $32 \times 5\% \approx 2$, there are 13 pixels in the ground-truth landmark circle of the backbone network. For landmark type k, we take $Ave\{M_c^k, 13\}$ as its coarse landmark location, where M_c^k is the k^{th} channel of M_c.

We resize I and M_c to 64×64. For landmark type k, we crop a 24×24 patch centered at the coarse landmark location from the concatenation of I and M_c^k, and we feed the patch into the k^{th} subnet of the branch networks to generate the fine map M_f^k. As the radius R_f is set to $64 \times 3\% \approx 2$, we take $Ave\{M_f^k, 13\}$ as the final location of landmark type k.

2) Unconstrained Setting: Given an unconstrained image, we construct an image pyramid of L levels by first resizing the image to make the length of the smaller side equal to 32 and gradually upsampling it with a scale factor of 1.16. The level number L can be dynamically adjusted based on the acceptable minimum face size. For example, we set L as 20 to locate the landmarks of the tiny faces in the AFW [8] dataset.

We further feed the images at different pyramid levels to the backbone network for generating multiple coarse heat maps and denote the k^{th} channel of the coarse heat maps at the l^{th} level as $M_{c,l}^k$. When the response value at location (x, y) of $M_{c,l}^k$ is higher than a given threshold, we assert that there is a 12×12 candidate region of landmark type k centered at that position. We denote this candidate region with a tuple $\{k, l, v, (x, y)\}$, where v is the response value at location (x,y).

A single landmark may be detected multiple times at a specific level or at different levels of the image pyramid. To reduce redundancy, for each landmark type, we first map all landmark candidate regions to the original image and then adopt non-maximum suppression (NMS) with an IOU threshold of 0.5 on these regions based on their response values. For a remaining landmark candidate region $\{k, l, v, (x, y)\}$, we crop its corresponding 12×12 heat map patch from $M_{c,l}^k$ and the RGB patch

from the image at the l^{th} level of the pyramid and further resize these two patches to 24×24 before feeding them into the k^{th} subnet of the branch networks to generate the fine heat map M_f^k. The final landmark location is computed by $Ave\{M_f^k, 13\}$.

IV. Experimental Results

A. Datasets

The existing public datasets of facial landmark localization are either too small and contain only hundreds of images or have very limited variation across different samples, e.g., most of the samples are near-frontal faces. These two situations greatly limit the performance of facial landmark localization under unconstrained settings. Therefore, we build a large-scale dataset called SYSU16K, which contains 7317 images (6317 for training and 1000 for validation) with 16K faces collected from the Internet. Each face is accurately annotated with 72 landmarks. With a large variation, the faces in our dataset exhibit various poses, expressions, illuminations and resolutions, and they may have severe occlusions. In addition, to train our proposed BB-FCN, we also randomly select 7542 natural images (6542 for training and 1000 for validation) without any faces from Pascal-VOC2012 as negative samples.

In our experiment, we evaluate our method on four public challenging datasets: LFPW [23], AFW [8], AFLW [16] and 300W [17]. There is no overlap among the training, validation and evaluation datasets.

AFLW: This dataset contains 21,080 faces in the wild. This dataset is very suitable for evaluating the performance of face alignment across a large range of poses. The selection of testing images from AFLW is as in [12], which randomly chooses 3000 faces, and 39% of them are non-frontal.

AFW: This dataset contains 205 images (468 faces) collected in the wild. Invisible landmarks are not annotated, and each face is annotated with at most 6 landmarks.

LFPW: This dataset contains 1,132 training images and 300 testing images. The images in this dataset are given in the form of URLs, and some image links are no longer valid. We can only download 811 training images and 230 testing images.

300W: The training set (3148 images) of this dataset is collected from the training sets of several exiting datasets, including LFPW (811), HELEN [62] (2000) and AFW (337). The full testing set is split into two subsets: (1) the common subset consists of the testing sets of LFPW (224) and HELEN (330), and (2) the challenging subset is composed of 135 images from IBUG [17]. All the images in this dataset are annotated with 68 facial landmarks.

B. Evaluation Metric

To evaluate the accuracy of facial landmark localization, we adopt the mean (position) error as the metric. For a specific type of landmark, the mean error is calculated as the mean distance between the detected landmarks of the given type in all testing images and their corresponding ground-truth positions, normalized with respect to the interocular distance. The (position) error of a single landmark is defined as follows:

$$err = \frac{\sqrt{(x-x')^2 + (y-y')^2}}{l} \times 100\%, \tag{7}$$

where (x, y) and (x', y') are the ground-truth and detected landmark locations, respectively, and l is the interocular distance. For the 300W dataset, the interocular distance is set to the Euclidean distance between the outer corners of two eyes, while for the other three landmark datasets, it is denoted as the Euclidean distance between the center points of the two eyes. In our experiments, we evaluate the mean error of every type of facial landmark and the average mean error over all landmark types, i.e., LE (left eye), RE (right eye), N (nose), LM (left mouth corner) and RM (right mouth corner), as well as A (average mean error of the five facial landmarks).

C. Performance Evaluation for Unconstrained Settings

Our BB-FCN is capable of dealing with facial images taken in unconstrained settings, e.g., the locations of facial regions and the number of faces in the image are unknown. In this setting, we use the recall-error curves to evaluate the performance of all comparative methods. A predictive facial landmark is considered to be correct if there exists a ground-truth landmark of the same type within the given position error. For a fixed number m (such as 15 or 30) of predictive landmarks, the recall rate (the fraction of ground-truth annotations covered by predictive landmarks) varies as the acceptable position error increases; thus, a recall-error curve can be obtained.

To the best of our knowledge, very few facial landmark localization methods have been evaluated in the context of landmark detection under unconstrained settings. For fairness, we have also implemented a regression-based method using an FCN with nine convolutional layers, which can be expressed as follows: $C(20, 5 \times 5 \times 3)$ - $C(20, 5 \times 5 \times 20)$ - MP - $C(30, 5 \times 5 \times 20)$ - $C(30, 5 \times 5 \times 30)$ - MP - $C(40, 5 \times 5 \times 30)$ - $C(40, 5 \times 5 \times 40)$ - $C(30, 2 \times 2 \times 40)$ - $C(30, 4 \times 4 \times 30)$ - $C(15, 1 \times 1 \times 30)$. With a training strategy similar to that of our backbone network, this regression-based network also takes a 32×32 image patch as input and generates a $15 \times 8 \times 8$ response map, each pixel of which corresponds to a 4×4 region of the input image. We formulate every three channels of the output response map as a group. Additionally, each pixel on a specific group indicates the probability of existence and the regressed two-dimensional location of the corresponding landmark type in a 4×4 region. During the testing phase, the same image pyramid is fed into the regression-based network for facial landmark inference.

We evaluate the performance of our BB-FCN and the regression-based deep model on the AFW dataset using an unconstrained setting. For those faces where one or both eyes are invisible, the interocular distances are set as 41.9% of the length of their annotated bounding boxes[2]. Figure 5 shows the recall-error curves of different types of landmarks, where the curves labeled "fine" and "coarse" illustrate the performance

[2]The average ratio between the interocular distances of the common faces and the length of their annotated bounding boxes is 41.9% on AFW.

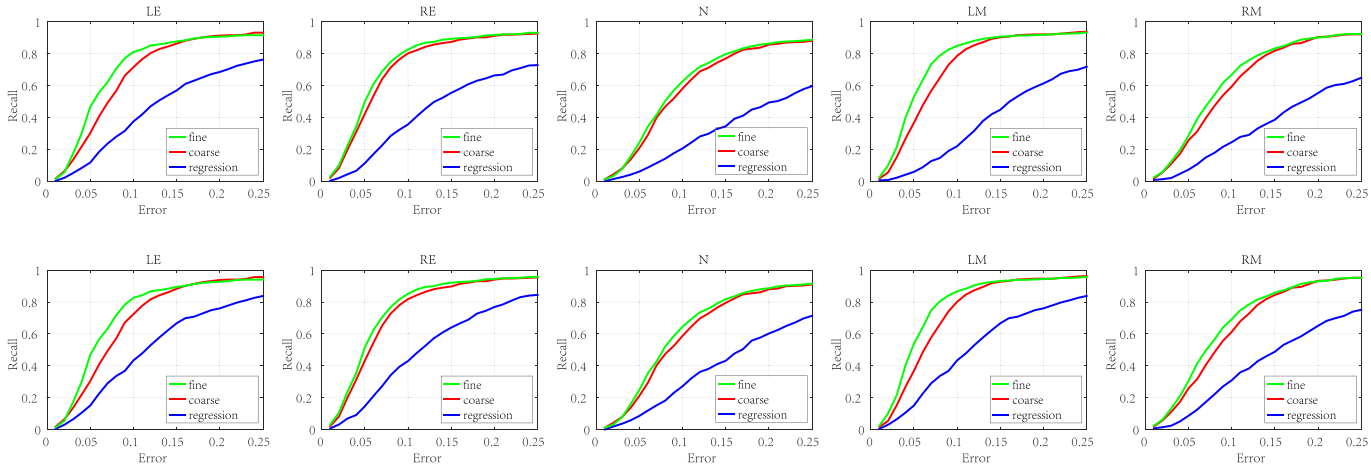


Fig. 5. The recall of landmarks on AFW in unconstrained settings. The curves labeled "fine" and "coarse" show the performance of models with and without branch networks, respectively. The curve labeled "regression" presents the performance of the regression network based on a single fully convolutional network. The top five figures demonstrate the recall performance when only 15 landmarks of each landmark type are predicted for each image, while the bottom five figures are the results with 30 predictive landmarks for each type of each image.

TABLE I
AVERAGE RECALLS OF THE COMPLETE BACKBONE-BRANCHES NETWORK AND THE BACKBONE NETWORK ALONE ON AFW IN UNCONSTRAINED SETTINGS. PE REFERS TO THE ACCEPTABLE POSITION ERROR

Model	15 landmarks		30 landmarks	
	PE=5%	PE=10%	PE=5%	PE=10%
backbone	31.1%	69.5%	31.5%	70.9%
full model	40.4%	75.6%	41.5%	77.6%

of our complete BB-FCN model and the single backbone network, respectively. The curve labeled "regression" indicates the performance of the above regression network based on a single FCN.

Our methods significantly outperform the regression network. With a prediction of 15 landmarks for each landmark type, the full model recalls 45% more landmarks than the regression network when the acceptable position error is set within 8% of the interocular distance. Given more predicted landmarks, our complete BB-FCN model can achieve higher landmark recalls. As the number of landmark predictions of each type increases to 30, the recalls of five landmarks within a position error of 25% of the interocular distance are 94.1%, 95.7%, 91.5%, 95.8% and 95.2%, respectively. Meanwhile, the full model performs much better than the backbone network alone. The average recalls of five landmarks are shown in Table I, which shows that the full model improves the recall rate by approximately 10% and 6% when the acceptable position error is set as 5% or 10%, respectively. As shown in Figure 6, our BB-FCN can generate high-quality heat maps and detect almost all the facial landmarks, even though some false positives exist. These false positives are some tiny and blurry regions (such as treetops and hands) that have rich texture or have similar shapes and colors as faces.

D. Performance Evaluation for Constrained Settings

In this setting, because the face bounding boxes are given, we can directly feed the face regions into our BB-FCN network to locate the facial landmarks. We will compare our method with state-of-the-art methods on the five landmark types and on dense landmark types.

1) Evaluation on Five Landmark Types: We compare our method with other state-of-the-art methods, i.e.,[3] robust cascaded pose regression (RCPR) [34], tree structured part model (TSPM) [8], Luxand face SDK,[4], explicit shape regression (ESR) [27], cascaded deformable shape model (CDM) [63], supervised descent method (SDM) [29], tasks-constrained deep convolutional network (TCDCN) [12], multitask cascaded convolutional networks (MTCNN) [64], recurrent attentive-refinement networks (RAR) [45], and unsupervised discovery (UD) [53].

On the AFW dataset, our average mean error over the five landmark types is 6.18%, which is an improvement over the performance of the state-of-the-art TDCN by 24.6%. On the AFLW dataset, our BB-FCN model achieves 6.28% average mean error, a 21.5% improvement over TDCN. Figure 7 and Table II demonstrate that our BB-FCN network outperforms all competing methods on the three datasets. The qualitative results presented in Figure 8 show that our method is robust under occlusions, exaggerated expressions and extreme illumination.

2) Evaluation on Dense Landmark Types: We can use our BB-FCN network for dense landmark prediction by simply extending the number of branches in the branch network. We evaluate our extended method on LFPW with 29 landmarks and on 300W with 68 landmarks. Because dense landmark prediction requires more facial details to distinguish landmarks with similar appearances, such as left-eyebrow-center-top and left-eyebrow-center-bottom, we enlarge the input images of BB-FCN to 64×64. Due to the differences between the landmark types of our collected dataset and LFPW, we fine-tune the network using the training set of LFPW. Moreover, for the 68

[3]Some results on AFW and AFLW are quoted from [12].
[4]Luxand face SDK: http://www.luxand.com/

Fig. 6. Qualitative facial landmark detection results in unconstrained settings. Our BB-FCN is capable of dealing with unconstrained facial images, even though the locations of facial regions and the number of faces in the image are unknown. Best viewed in color with zoom.

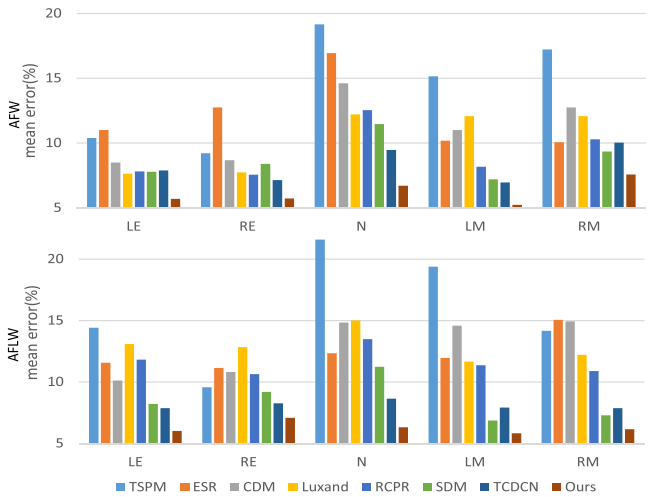


Fig. 7. Comparisons with state-of-the-art methods on two public datasets. The top row shows the corresponding results on AFW, and the bottom row shows the corresponding results on AFLW. The average mean errors of all considered methods are summarized in Table II.

TABLE II
AVERAGE MEAN ERRORS OF OUR METHOD AND OF ALL OTHER COMPETING METHODS ON AFW AND AFLW

Dataset	AFW	AFLW
TSPM	14.31	15.9
ESR	12.2	13
CMD	11.1	13.1
Luxand	10.4	12.4
RCRR	9.3	11.6
SDM	8.8	8.5
TCDCN	8.2	8.0
RAR	-	7.23
MTCNN	-	6.9
UD	-	6.58
Ours	6.18	6.28

landmark types, we train our network from scratch with the training set of the 300W dataset.

We compare our method with other state-of-the-art methods on the LFPW dataset. The other methods include consensus of exemplars (CE) [23], explicit shape regression (ESR) [27] and ensemble of regression trees (ERT) [28]. Table IV shows that our BB-FCN achieves 3.35% average mean error, outperforming the other three state-of-the-art methods.

We also compare the performance of our proposed method with the results of other state-of-the-art methods on the 300W testing set with 68 landmarks. The first class of compared methods are cascaded regression-based models, including TSPM, RCPR, SDM, ESR, LBF [30] and CFSS [65]. The second class are deep-learning-based methods, including TCDCN, 3DDFA [47], CFAN [38], RAR [45], Pose-Invariant [66], RDR [67], Two-Stage$_{OD}$ [68], and RCN^{+} [55]. As shown in Table III, our proposed method significantly outperforms all the other state-of-the-art methods across all different testing sets; specifically, our complete model lowers the average mean error achieved by the best-performing existing algorithm (RCN^{+}) by 8.3%, 3.6% and 6.9% on the common set, the challenging set and the full set, respectively. Figure 9 presents some example results of our proposed pixel-labeling method for dense landmark prediction.

E. Ablation Study

Our proposed BB-FCN is composed of two components: the backbone network and the branch networks. To show the effectiveness and necessity of these two components, we compare the landmark prediction results produced by the single backbone network with those of the complete BB-FCN network. As shown in Table V, the average mean error on AFLW is decreased from 8.31% to 6.28%, with an approximately 24.4% relative improvement, after the branch networks are added to perform landmark refinement. The quantitative comparison shown in Figure 10 further demonstrates that the prediction error of every type of

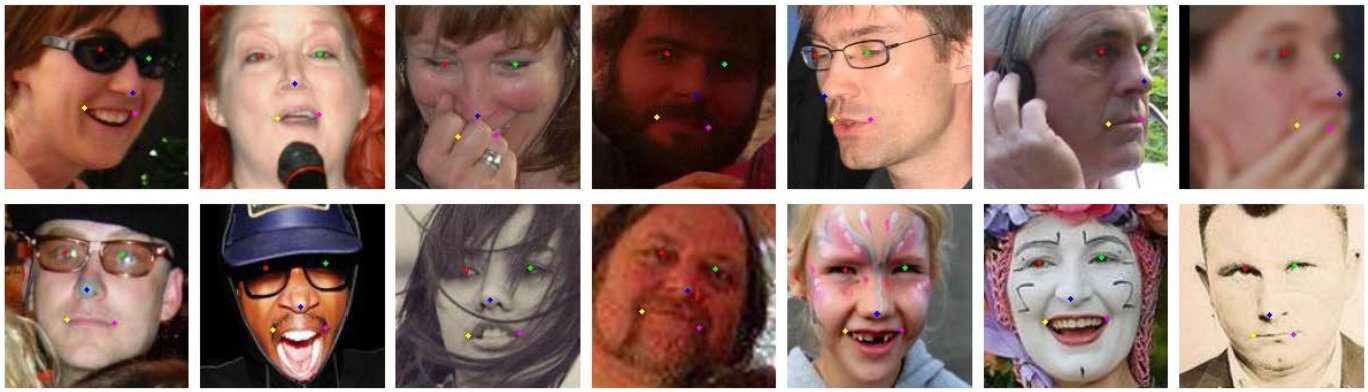


Fig. 8. Qualitative facial landmark localization results of our method. The first row shows the result on AFW, while the second row shows the result on AFLW. Our method is robust under occlusions, exaggerated expressions and extreme illumination.

TABLE III
AVERAGE MEAN ERRORS OF LANDMARK DETECTION ON THE 300W DATASET

Methods	Common Set	Challenging Set	Full Set
TSPM	8.22	18.33	10.22
RCPR	6.18	17.26	8.35
SDM	5.57	15.40	7.50
ESR	5.28	17.00	7.58
LBF	4.95	11.98	6.32
CFSS	4.73	9.98	5.76
CFAN	5.50	16.78	7.69
3DDFA	6.15	10.59	7.01
3DDFA+SDM	5.53	9.56	6.31
TCDCN	4.8	8.6	5.54
RAR	4.12	8.35	4.94
Pose-Invariant	5.43	9.88	6.30
RDR	5.03	8.95	5.80
Two-Stage$_{OD}$	4.36	7.56	4.99
RCN^{+}	4.20	7.78	4.90
Ours	**3.85**	**7.50**	**4.56**

TABLE IV
AVERAGE MEAN ERRORS OF OUR METHOD AND ALL OTHER COMPETING METHODS ON LFPW

Methods	CE	ESR	ERT	Ours
Error	4.00	3.43	3.80	3.35

facial landmark enjoys a varying degree of reduction on LFPW. Figure 11 shows the visual improvements achieved with the branch networks over the single backbone network. As shown, the output heat maps of the branch networks are more compact and precise than those of the backbone network, which can well explain the better performance of branch networks.

F. The Effectiveness of Face Proposal Generation

In this experiment, we demonstrate the effectiveness of our landmark prediction network in face proposal generation. A predicted facial landmark typically indicates the existence of a face; therefore, we can generate face proposals from the response heat map of the BB-FCN. For a type-k predicted facial landmark at level l, we generate a 32×32 face candidate window centered at the landmark location from the RGB image at the l^{th} level of the pyramid. We then apply NMS to face proposals generated using each type of landmark. After fine-tuning the location and edge length of face proposals with Net-12 (the first network of cascade CNN [69]), we apply NMS to all face proposals again.

We compare our method with three generic object proposal generators [70]–[72] and a face-specific proposal generator, Faceness [73], on FDDB. For a fair comparison, following [73], we also transform the original ground-truth ellipses in FDDB into minimal rectangular bounding boxes. Table VI[5] shows that our method achieves high recalls using a very small number of face proposals due to the accuracy of landmark localization. Our method can detect 72.8% of faces using only two proposals per image and 81.2% of faces using three proposals per image on FDDB. It detects 91.5% of faces when at most 20 face proposals are generated from each image. With a similar proposal generation strategy as our method, Faceness [73] utilizes facial attributes to calculate the facial part response maps and then generates the region proposal from these response maps. Compared with [73], our method can generate more accurate landmark (part) response maps by explicitly locating the facial landmarks, and it is more robust to overcome partial occlusions and head pose variations.

G. Evaluation on Face Detection Performance

Our BB-FCN network can locate various landmarks in unconstrained settings and generate high-quality face proposals, which can enhance the performance of existing face detectors, such as cascade CNN [69], particularly under severe occlusions and large pose variation. Cascade CNN is one of the up-to-date fast face detectors. It relies on six cascaded convolutional neural networks to locate faces in an image. We retrain this detector using our collected landmark dataset and Pascal VOC 2012, and we achieve similar performance on FDDB. We replace the original face proposals used by cascade CNN with our landmark-based proposals. All other parts of the method remain the same. The experimental results indicate that the modified cascade CNN achieves state-of-the-art performance on two public face detection benchmarks: FDDB and AFW.

1) FDDB: As a large-scale face detection benchmark, FDDB contains 5,171 annotated faces in 2,845 images. It uses

[5]The results from the compared methods are quoted from [73].

Fig. 9. Qualitative facial landmark localization results of our method on the 300W dataset. The first row shows the result on the common set, while the second row demonstrates the result on the challenging set.

TABLE V
AVERAGE MEAN ERRORS OF THE COMPLETE BACKBONE-BRANCHES NETWORK AND THE BACKBONE NETWORK ON AFW AND AFLW

landmark type	AFW		AFLW	
	backbone	full model	backbone	full model
LE	7.02	5.69	9.46	6.02
RE	6.79	5.72	8,60	7.08
N	8.35	6.71	8.39	6.31
LM	7.11	5.22	7.40	5.83
RM	7.98	7.58	7.73	6.15
A	7.45	6.18	8.31	6.28

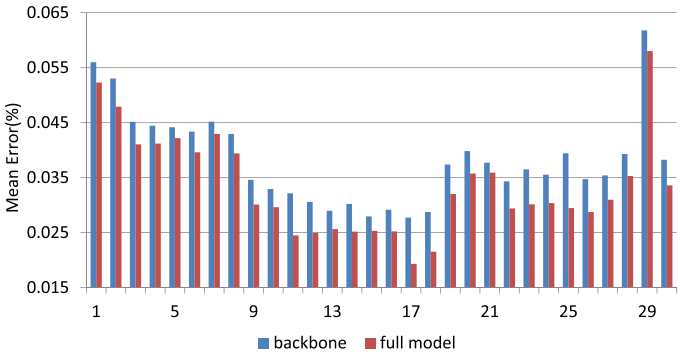


Fig. 10. Performance evaluation of the complete backbone-branches network and the backbone network alone on LFPW. The mean error of every type of landmark is decreased to a certain degree when the branch networks are used. The 30th column is the average mean error.

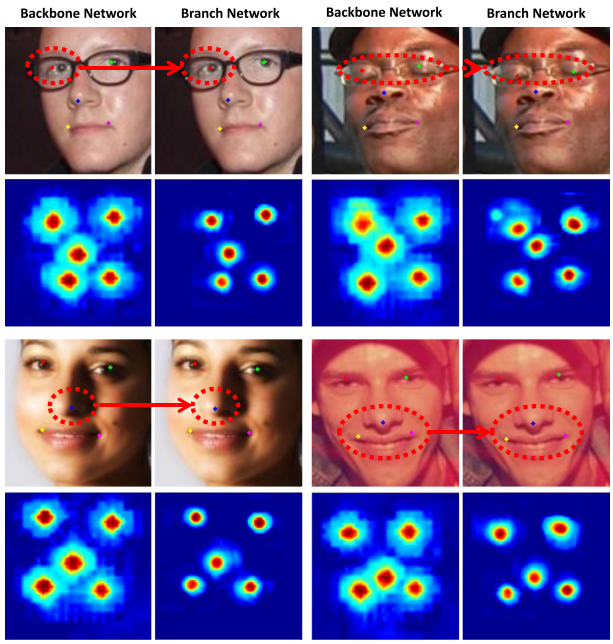


Fig. 11. Examples of improvements made by the branch networks. The response heat maps of the branch networks are more compact and precise. Best viewed in color.

TABLE VI
NUMBER OF PROPOSALS NEEDED FOR DIFFERENT RECALL RATES ON FDDB

Proposal method	75%	80%	85%	90%
EdgeBox	132	214	326	600
MCG	191	292	453	941
Selective Search	153	228	366	641
EdgeBox+Faceness	21	47	99	288
MCG+Faceness	13	23	55	158
Selective Search+Faceness	24	41	91	237
Ours	>2	<3	5	9

elliptic face annotations and defines two types of evaluations: the discontinuous score and continuous score. We use the discontinuous score evaluation, which counts the number of detected faces versus the number of false alarms. A detected bounding box is taken as the true positive only if the IoU between this bounding box and the bounding box of a ground-truth face is above 0.5. We uniformly enlarge our square bounding boxes vertically by 25% to better approximate elliptic annotations in FDDB.

As shown in Figure 12, face proposals defined by different landmark types exhibit different levels of effectiveness in face detection. The nose landmark achieves the best performance among all landmark types. Using face proposals defined by all five landmark types significantly improves the performance achieved with individual landmark types.

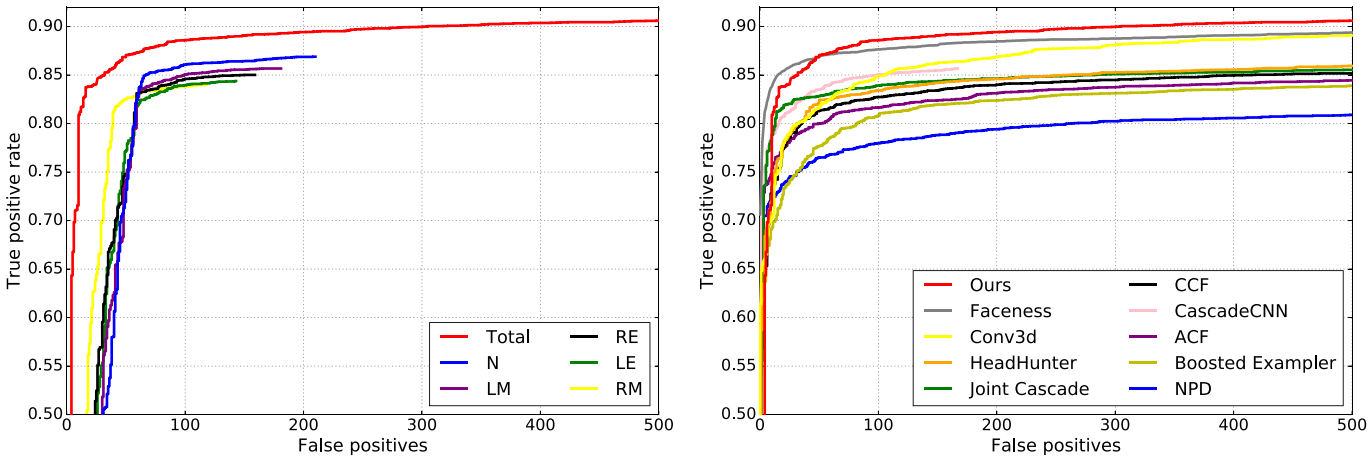


Fig. 12. Left: Face proposals induced by different landmark types exhibit different levels of effectiveness in face detection. Using face proposals induced by all five landmark types significantly improves the performance achieved with individual landmark types. Right: On the FDDB dataset, we compare our method against other state-of-the-art methods. When the number of false positives is fixed at 350, the recall achieved with our method is 90.17%, which is higher than all other methods.

We compare our method with nine recently published state-of-the-art methods on the FDDB dataset. These methods include cascade CNN [69], Faceness [73], CCF [74], Conv3d [75], HeadHunter [76], joint cascade [77], boosted exemplar [78], ACF [79] and NDP [78]. Figure 12 shows that our method outperforms all nine state-of-the-art methods by a considerable margin. When the number of false positives is fixed at 167, our method achieves a significant margin of 3.51% in recall rate over the baseline cascade-CNN [69]. When the number of false positives is fixed at 350, our method achieves a 90.17% recall rate, which is higher than the 88.92% recall rate achieved by Faceness [73]. When the number of false positives increases to 500, our method obtains a recall rate of 90.6% with at most 20 face proposals per images. In contrast, when trained with approximately 83K face images, joint training cascade CNN [80] generates nearly 1000 proposals on average before applying the MNS and only obtains a recall of 88.2% with 1000 false positives. Recently, SAFD [81] trained their network with 350K private face images and obtained a recall of 93.8% with 1000 false positives. However, our method can achieve competitive performance with only 16K face images in our SYSU16K dataset.

2) AFW: We adopt the precision-recall protocol when performing evaluation on the AFW dataset. We compare our method with Faceness [73], HeadHunter [76], structured models [82], SquareChnFtrs-5[76], Shen et al. [83], TSM [8], Face.com, Face++ and Picasa. As shown in Figure 13, with an average precision of 97.46%, the performance of our detector is comparable to that of other state-of-the-art techniques.

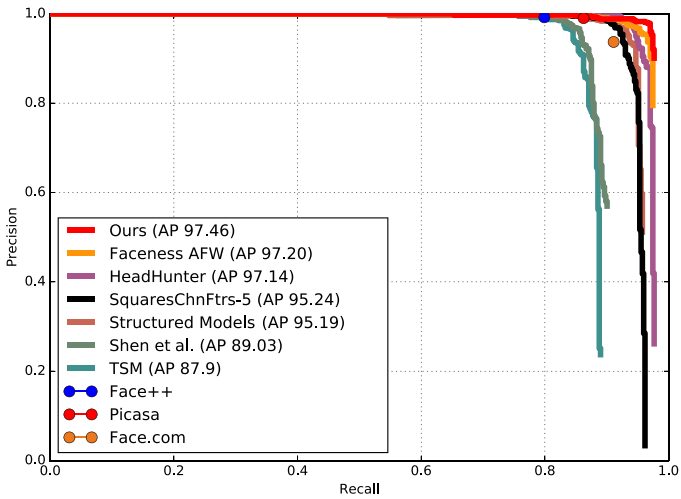


Fig. 13. Precision-recall curves of 10 face detection methods on the AFW dataset. The performance of our face detector is comparable to that of other state-of-the-art techniques.

H. Limitations

In this section, we present failure cases of our BB-FCN network. In our experiments, we found that BB-FCN occasionally generates results that do not conform to the normal spatial layout of human facial landmarks, as shown in Figure 14(a). The main reason for this phenomenon is the lack of constraints on relative landmark positions in the loss function. Second, BB-FCN fails to highlight facial landmarks in blurry images, as shown in Figure 14(b). This negatively impacts the performance of our face proposal method on FDDB, which contains many blurry faces.

I. Runtime Efficiency

One of the most important characteristics of our landmark and face detectors is the efficiency. Our method achieves practical runtime efficiency via a coarse-to-fine pipeline. Table VII shows the running times of several deep models for five facial landmark detection under constrained settings. Among these models, TCDCN requires 18 ms to process a facial image on an Intel Core i5 CPU, which is 7 times faster than CDCN [13]. CFAN [38] costs 30 ms to run multiple autoencoders. Our method only needs 9 ms on an Intel Core i5 2.80 GHz CPU and 1.8 ms on an NVIDIA Titan X GPU. For the localization of

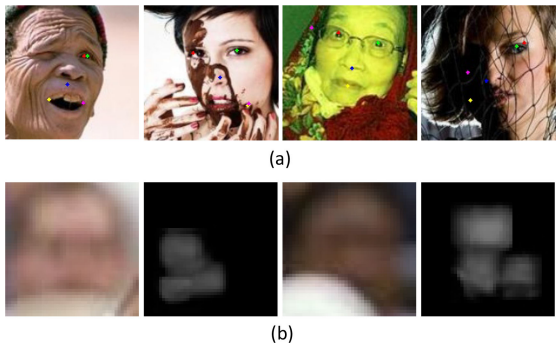


Fig. 14. Failure cases of our BB-FCN network. (a) Incorrect landmark prediction results that violate the normal spatial layout of human facial landmarks. (b) Two blurry faces from FDDB and their response heat maps.

TABLE VII
COMPARISON OF RUNNING TIMES ON CPU AMONG DEEP MODELS FOR FIVE FACIAL LANDMARK DETECTION

Methods	Time(per face)
CDCN	120 ms
CFAN	30 ms
TCDCN	18 ms
Ours	9 ms

68 landmarks, our method costs 10 ms to process a face region on the same GPU.

For the unconstrained setting, to locate the landmarks of the tiny faces for a high recall rate, we build a 20-level image pyramid on the AFW and FDDB datasets, and our landmark network runs at approximately 6 PFS on the same GPU. However, the level number of the image pyramid can be dynamically adjusted based on the acceptable minimum face size. For example, to locate the landmarks of faces with sizes larger than 80×80 from 640×480 VGA images, we only need to build an image pyramid with 7 levels. In this case, our landmark networks can run at 30 FPS, while our face detection pipeline can run at approximately 20 FPS on the same GPU thanks to our efficient proposal generator and the cascade CNN detector. For comparison, Shen *et al.* [83] process a 1280-pixel wide image in less than 10 seconds and DP2MFD [84] runs at 0.285 FPS on an Nvidia Tesla K20, while the ResNet101-based detector proposed by HR [85] runs at 3.1 FPS on 720p resolution. With a similar speed as our network, Faceness [73] can process a VGA image within 50 ms on a Titan Black GPU, but their performance is worse than ours.

V. CONCLUSION

In this paper, we have presented a novel cascaded backbone-branches fully-convolutional network (BB-FCN) that progressively produces response maps of facial landmarks in an end-to-end manner. Our extensive experiments demonstrate that BB-FCN achieves very promising results on both traditional benchmarks with a controlled setting and on cluttered, real-world scenes. When exploiting our facial landmark localization results in R-CNN-based face detection, we have observed a significant increase in both accuracy and efficiency. In the future, we will integrate our BB-FCN model with object recognition and detection systems where accurate part-based localization can be helpful in improving object detection performance.

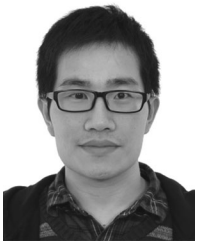
REFERENCES

[1] P. Luo, X. Wang, and X. Tang, "A deep sum-product architecture for robust facial attributes analysis," in *Proc. IEEE Int. Conf. Comput. Vision*, 2013, pp. 2864–2871.
[2] C. Lu and X. Tang, "Surpassing human-level face verification performance on LFW with gaussian face," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 3811–3819.
[3] L. Liu *et al.*, "Deep aging face verification with large gaps," *IEEE Trans. Multimedia*, vol. 18, no. 1, pp. 64–75, Jan. 2016.
[4] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep learning identity-preserving face space," in *Proc. IEEE Int. Conf. Comput. Vision*, 2013, pp. 113–120.
[5] C. Ding and D. Tao, "Robust face recognition via multimodal deep face representation," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2049–2058, Nov. 2015.
[6] Y. Li, L. Liu, L. Lin, and Q. Wang, "Face recognition by coarse-to-fine landmark regression with application to ATM surveillance," in *Proc. Chin. Conf. Comput. Vision*, Springer, 2017, pp. 62–73.
[7] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2001, vol. 1, pp. I–511.
[8] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2012, pp. 2879–2886.
[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 580–587.
[10] L. Liu, H. Wang, G. Li, W. Ouyang, and L. Lin, "Crowd counting using deep recurrent spatial-aware network," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 849–855.
[11] L. Liu *et al.*, "Attentive crowd flow machines," in *Proc. Comput. Vision Pattern Recognit.*, 2018, pp. 1553–1561.
[12] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. Eur. Conf. Comput. Vision*, Springer, 2014, pp. 94–108.
[13] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 3476–3483.
[14] R. Weng, J. Lu, Y.-P. Tan, and J. Zhou, "Learning cascaded deep auto-encoder networks for face alignment," *IEEE Trans. Multimedia*, vol. 18, no. 10, pp. 2066–2078, Oct. 2016.
[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3431–3440.
[16] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Proc. IEEE Int. Conf. Comput. Vision Workshops*, 2011, pp. 2144–2151.
[17] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proc. IEEE Int. Conf. Comput. Vision Workshops*, 2013, pp. 397–403.
[18] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.
[19] P. Sauer, T. F. Cootes, and C. J. Taylor, "Accurate regression procedures for active appearance models," in *Proc. Brit. Mach. Vision Conf.*, 2011, pp. 1–11.
[20] P. A. Tresadern, P. Sauer, and T. F. Cootes, "Additive update predictors in active appearance models," in *Proc. Brit. Mach. Vision Conf.*, Citeseer, 2010, vol. 2., p. 4.
[21] G. Tzimiropoulos and M. Pantic, "Optimization problems for fast AAM fitting in-the-wild," in *Proc. IEEE Int. Conf. Comput. Vision*, 2013, pp. 593–600.
[22] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *Int. J. Comput. Vision*, vol. 91, no. 2, pp. 200–215, 2011.
[23] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2930–2940, Dec. 2013.

[24] L. Liang, R. Xiao, F. Wen, and J. Sun, "Face alignment via component-based discriminative search," in *Proc. Eur. Conf. Comput. Vision*, Springer, 2008, pp. 72–85.
[25] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool, "Real-time facial feature detection using conditional regression forests," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2012, pp. 2578–2585.
[26] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2010, pp. 2729–2736.
[27] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *Int. J. Comput. Vision*, vol. 107, no. 2, pp. 177–190, 2014.
[28] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 1867–1874.
[29] X. Xiong and F. Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 532–539.
[30] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 1685–1692.
[31] S. Zhu, C. Li, C.-C. Loy, and X. Tang, "Unconstrained face alignment via cascaded compositional learning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 3409–3417.
[32] O. Tuzel, T. K. Marks, and S. Tambe, "Robust face alignment using a mixture of invariant experts," in *Proc. Eur. Conf. Comput. Vision*, Springer, 2016, pp. 825–841.
[33] X. Fan, R. Liu, Z. Luo, Y. Li, and Y. Feng, "Explicit shape regression with characteristic number for facial landmark localization," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 567–579, Mar. 2018.
[34] X. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proc. IEEE Int. Conf. Comput. Vision*, 2013, pp. 1513–1520.
[35] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade," in *Proc. IEEE Int. Conf. Comput. Vision Workshops*, 2013, pp. 386–391.
[36] H. Liu, D. Kong, S. Wang, and B. Yin, "Sparse pose regression via componentwise clustering feature point representation," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1233–1244, Jul. 2016.
[37] T. Zhang *et al.*, "A deep neural network-driven feature learning method for multi-view facial expression recognition," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2528–2536, Dec. 2016.
[38] J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment," in *Proc. Eur. Conf. Comput. Vision*, Springer, 2014, pp. 1–16.
[39] J. Zhang, M. Kan, S. Shan, and X. Chen, "Occlusion-free face alignment: Deep regression networks coupled with de-corrupt autoencoders," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 3428–3437.
[40] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, 2017.
[41] D. Merget, M. Rock, and G. Rigoll, "Robust facial landmark detection via a fully-convolutional local-global context network," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 781–790.
[42] A. Bulat and G. Tzimiropoulos, "Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans," in *Proc. Comput. Vision Pattern Recognit.*, 2018, pp. 109–117.
[43] Z. Tang *et al.*, "Quantized densely connected U-nets for efficient landmark localization," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 339–354.
[44] X. Peng, R. S. Feris, X. Wang, and D. N. Metaxas, "A recurrent encoder-decoder network for sequential face alignment," in *Proc. Eur. Conf. Comput. Vision*, Springer, 2016, pp. 38–56.
[45] S. Xiao *et al.*, "Robust facial landmark detection via recurrent attentive-refinement networks," in *Proc. Eur. Conf. Comput. Vision*, Springer, 2016, pp. 57–72.
[46] 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 *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 4177–4187.
[47] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 146–155.
[48] A. Jourabloo and X. Liu, "Large-pose face alignment via CNN-based dense 3d model fitting," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 4188–4196.
[49] F. Liu, D. Zeng, Q. Zhao, and X. Liu, "Joint face alignment and 3d face reconstruction," in *Proc. Eur. Conf. Comput. Vision*. Springer, 2016, pp. 545–560.
[50] 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 *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, vol. 1, no. 2, pp. 1021–1030.
[51] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3d face reconstruction and dense alignment with position map regression network," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 534–551.
[52] X. Dong *et al.*, "Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 360–368.
[53] Y. Zhang *et al.*, "Unsupervised discovery of object landmarks as structural representations," in *Proc. Comput. Vision Pattern Recognit.*, 2018, pp. 2694–2703.
[54] X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Style aggregated network for facial landmark detection," in *Proc. Comput. Vision Pattern Recognit.*, 2018, vol. 2, pp. 379–388.
[55] S. Honari *et al.*, "Improving landmark localization with semi-supervised learning," in *Proc. Comput. Vision Pattern Recognit.*, 2018, pp. 1546–1555.
[56] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 1395–1403.
[57] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep CNN features," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5012–5024, Nov. 2016.
[58] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li, "Disc: Deep image saliency computing via progressive representation learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1135–1149, Jun. 2016.
[59] G. Li and Y. Yu, "Contrast-oriented deep neural networks for salient object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6038–6051, Dec. 2018.
[60] Z. Yan *et al.*, "HD-CNN: Hierarchical deep convolutional neural networks for large scale visual recognition," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 2740–2748.
[61] M. Everingham *et al.*, "The pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vision*, vol. 111, no. 1, pp. 98–136, 2015.
[62] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. Huang, "Interactive facial feature localization," in *Proc. Eur. Conf. Comput. Vision*, 2012, pp. 679–692.
[63] X. Yu, J. Huang, S. Zhang, W. Yan, and D. Metaxas, "Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model," in *Proc. IEEE Int. Conf. Comput. Vision*, 2013, pp. 1944–1951.
[64] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[65] S. Zhu, C. Li, C. Change Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 4998–5006.
[66] A. Jourabloo, M. Ye, X. Liu, and L. Ren, "Pose-invariant face alignment with a single CNN," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 3200–3209.
[67] S. Xiao *et al.*, "Recurrent 3d-2d dual learning for large-pose facial landmark detection," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 1642–1651.
[68] J.-J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou, "A deep regression architecture with two-stage re-initialization for high performance facial landmark detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 3691–3700.
[69] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 5325–5334.
[70] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vision*, Springer, 2014, pp. 391–405.
[71] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vision*, vol. 104, no. 2, pp. 154–171, 2013.
[72] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 328–335.
[73] S. Yang, P. Luo, C. C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 3676–3684.

[74] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Convolutional channel features," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 82–90.
[75] Y. Li *et al.*, "Face detection with end-to-end integration of a convnet and a 3d model," in *Proc. Eur. Conf. Comput. Vision*, Cham, Switzerland, 2016, pp. 420–436.
[76] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *Proc. Eur. Conf. Comput. Vision*, Springer, 2014, pp. 720–735.
[77] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, "Joint cascade face detection and alignment," in *Proc. Eur. Conf. Comput. Vision*, Springer, 2014, pp. 109–122.
[78] H. Li, Z. Lin, J. Brandt, X. Shen, and G. Hua, "Efficient boosted exemplar-based face detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 1843–1850.
[79] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Aggregate channel features for multi-view face detection," in *Proc. IEEE Int. Joint Conf. Biometrics*, 2014, pp. 1–8.
[80] H. Qin, J. Yan, X. Li, and X. Hu, "Joint training of cascaded cnn for face detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 3456–3465.
[81] Z. Hao *et al.*, "Scale-aware face detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, vol. 3, pp. 1913–1922.
[82] J. Yan, X. Zhang, Z. Lei, and S. Z. Li, "Face detection by structural models," *Image Vision Comput.*, vol. 32, no. 10, pp. 790–799, 2014.
[83] X. Shen, Z. Lin, J. Brandt, and Y. Wu, "Detecting and aligning faces by image retrieval," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 3460–3467.
[84] R. Ranjan, V. M. Patel, and R. Chellappa, "A deep pyramid deformable part model for face detection," in *Proc. IEEE 7th Int. Conf. Biometrics Theory, Appl. Syst.*, 2015, pp. 1–8.
[85] P. Hu and D. Ramanan, "Finding tiny faces," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 1522–1530.

Lingbo Liu received the B.E. degree in 2015 from the School of Software, Sun Yat-sen University, Guangzhou, China, where he is currently working toward the Ph.D. degree in computer science at the School of Data and Computer Science. His current research interests include computer vision, big data analysis, and intelligent transportation systems.



Guanbin Li (M'16) received the Ph.D. degree from the University of Hong Kong, Hong Kong, in 2016. He is currently a Research Associate Professor with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He has authored and coauthored more than 30 papers in top-tier academic journals and conferences. His current research interests include computer vision, image processing, and deep learning. He was a recipient of the Hong Kong Postgraduate Fellowship. He is an Area Chair for the conference of VISAPP. He is a Reviewer for numerous academic journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, TC, Conference on Computer Vision and Pattern Recognition, and International Joint Conferences on Artificial Intelligence.

Yuan Xie received the B.E. degree from the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China, where she is currently working toward the M.E. degree. Her research interests include facial alignment and salient object detection.

Yizhou Yu (M'10–SM'12–F'19) received the Ph.D. degree from the University of California at Berkeley, Berkeley, CA, USA, in 2000. He is a Professor with The University of Hong Kong, and was a faculty member with the University of Illinois at Urbana-Champaign for twelve years. His current research interests include computer vision, deep learning, biomedical data analysis, computational visual media, and geometric computing. He is a recipient of 2002 U.S. National Science Foundation CAREER Award, 2007 NNSF China Overseas Distinguished Young Investigator Award, and ACCV 2018 Best Application Paper Award. He was the Editorial Board of IET Computer Vision, The Visual Computer, and IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS. He was on the program committee of many leading international conferences, including SIGGRAPH, SIGGRAPH Asia, and the International Conference on Computer Vision.

Qing Wang received the Ph.D. degree from the School of Data and Computer Science, Sun Yat-sen University, where he is an Associate Professor and an Experienced Researcher on human computer interaction, software engineering, software architecture, computer vision, and artificial intelligence. He received the Best Paper Honorable Mention Award in ACM CHI 2010, Google Research/Education Award in 2010, and Google Faculty Award in 2011.

Liang Lin (M'09–SM'15) is a Full Professor of Sun Yat-sen University, Guangzhou, China. He is an Excellent Young Scientist of the National Natural Science Foundation of China. From 2008 to 2010, he was a Postdoctoral Fellow with the University of California, Los Angeles. From 2014 to 2015, as a senior visiting scholar, he was with The Hong Kong Polytechnic University and The Chinese University of Hong Kong. He currently leads the SenseTime R&D teams to develop cutting-edge and deliverable solutions on computer vision, data analysis and mining, and intelligent robotic systems. He has authored and coauthored more than 100 papers in top-tier academic journals and conferences. He is an Associate Editor of the IEEE TRANSACTIONS HUMAN-MACHINE SYSTEMS, The Visual Computer and Neurocomputing. He was area/session chairs for numerous conferences, such as the Institute for Computational and Mathematical Engineering, Asian Conference on Computer Vision, ICMR. He was the recipient of the Best Paper Runners-Up Award in ACM NPAR 2010, the Google Faculty Award in 2012, the Best Paper Diamond Award in IEEE International Conference on Multimedia and Expo 2017, and the Hong Kong Scholars Award in 2014. He is a Fellow of IET.