Facial Landmark Localization in the Wild by Backbone-Branches Representation Learning

Lingbo Liu[1], Guanbin Li[1]*, Yuan Xie[1], Yizhou Yu[2], Liang Lin[1]
[1] School of Data and Computer Science, Sun Yat-sen University, GuangZhou, China
[2] Department of Computer Science, The University of Hong Kong, HongKong
liulingb@mail2.sysu.edu.cn, liguanbin@mail.sysu.edu.cn, xiey39@mail2.sysu.edu.cn
yizhouy@acm.org, linliang@ieee.org

***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 pre-processing. It follows a coarse-to-fine cascaded pipeline, which consists of a backbone network for roughly detecting the locations of all facial landmarks and one branch network for each type of detected landmarks for further refining their locations. 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.**

***Index Terms*—facial landmark, backbone-branches, unconstrained settings**

I. Introduction

Facial landmark localization aims at automatically predicting key point positions in facial image regions. It is an essential component in many face-related applications, such as face verification [1] and face recognition [2], [3]. Though tremendous effort has been made on this topic, its performance is still far from being perfect, especially on facial regions with severe occlusion or extreme head poses.

Most of existing approaches for facial landmark localization have been developed for a controlled setting, e.g., the facial regions are detected in a pre-processing step. This setting has drawbacks when we deal 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, with multiple faces, extreme head poses and occlusion (see Figure 1). More specifically, we keep in mind the following issues when developing our algorithm.

- Faces may have large appearance and structure variations in an unconstrained setting due to diverse viewing conditions, rich facial expressions and large pose changes. Therefore, traditional global models may not work well as the usual assumptions (e.g., certain spatial layouts) may not hold in such an environment.

*Corresponding author is Guanbin Li.

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

- The search space of facial landmarks is quite large under the circumstance that the number and the size-scale of person faces are both unknown. Thus it is quite infeasible and inefficient to handle our task by existing models with exhaustive image pyramid sliding-window searching.

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

With recent advances in deep learning techniques, deep convolutional neural network models have demonstrated significant progress in various tasks [4]–[8]. More recently, Long et al. [9] proposed a fully convolutional network (FCN) for pixel labeling, which takes an input image with an arbitrary size and produces a dense label map in the same resolution. It 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 a FCN in our

978-1-5386-5321-0/18/$31.00 ©2018 IEEE

task, i.e., pixelwise facial landmark prediction. Nevertheless, a specialized architecture is required as our task demands 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 refined by observing local regions. To this end, we introduce a novel architecture of fully convolutional neural networks that naturally follows this coarse-to-fine pipeline. Specifically, our architecture contains one backbone network and several branches each corresponding to one landmark type. For computational efficiency, the backbone network is designed to be a fully convolutional network which takes a whole low-resolution image as its input and rapidly generates an initial multi-channel heat map with each channel predicting the locations of a specific landmark. Given the initial heat map, we obtain landmark proposals as local maxima within each channel. We crop a region centered at every landmark proposal from both the original input image and the corresponding channel of the response map, and these cropped regions are stacked together and fed to a branch network for a fine and accurate localization. As fully connected layers are not used in either networks, we call our architecture as the cascaded Backbone-Branches Fully Convolutional Network (BB-FCN). Thanks to the specially 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 facial landmarks in unconstrained settings in real time.

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 each for refining the predictions of one specific type of landmarks.
- We extensively evaluate BB-FCN on several standard benchmarks (e.g., AFW [10], AFLW [11]), and our experiments show that BB-FCN achieves superior performance in comparison to other state-of-the-art methods under both the constrained (i.e., with face detections) and the unconstrained settings.

II. Related Work

Facial landmark localization has long been attempted in computer vision. And a large number of approaches have been proposed, which can be divided into two categories, template fitting methods and regression based methods.

Template fitting methods build face templates to fit input face appearance [12]–[15]. A representative work is the active appearance model (AAM) [16], which attempts to estimate model parameters through minimizing the residual between the holistic appearance and an appearance model. Instead of holistic representations, a constrained local model (CLM) [17] learns an independent local detector for each facial keypoint, and a shape model for capturing valid facial deformations. Improved versions of CLM [18] primarily differ from each other in terms of local detectors. These methods are usually superior to the holistic methods due to the robustness of patch detectors against illumination variations and occlusion.

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.* [19] estimated the absolute coordinates of facial landmarks directly from an ensemble of conditional regression trees trained on facial appearances. Cascaded regression models [20]–[24] formulate shape estimation as a regression problem and make predictions in a cascaded manner. They 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. All these methods assume that an initial shape is given in some form, e.g., a mean shape [21]. However, this assumption is too strict and may lead to poor performance on faces with large pose variations.

Recently, convolutional neural networks also have been successfully applied to facial landmark estimation [25]. Zhou *et al.* [26] proposed a four-level cascaded regression model based on CNNs, which sequentially predict landmark coordinates. Zhang *et al.* [27] proposed a new coarse-to-fine DAE pipeline to progressively refine facial landmark locations. RNN-based models [28]–[30] formulate facial landmarks detection as a sequential refine process in an end-to-end manner. 3D face models [31]–[33] are also utilized to improve the facial landmarks detection. Though these methods have achieved remarkable performance, all of them were developed for a controlled setting, which requires an image region bounding 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 restrict the flexibility in practice due to the fixed form of the parameterization. Such trained models struggle in unconstrained settings (e.g., the unknown number of faces in an image). In contrast, our approach produces pixel-wise response maps, making it very flexible in localizing facial landmarks in the Wild as well as 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 at locating all facial landmarks in this 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 the image I, where x_i^k, 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. Here, the five landmark types are respectively the left eye

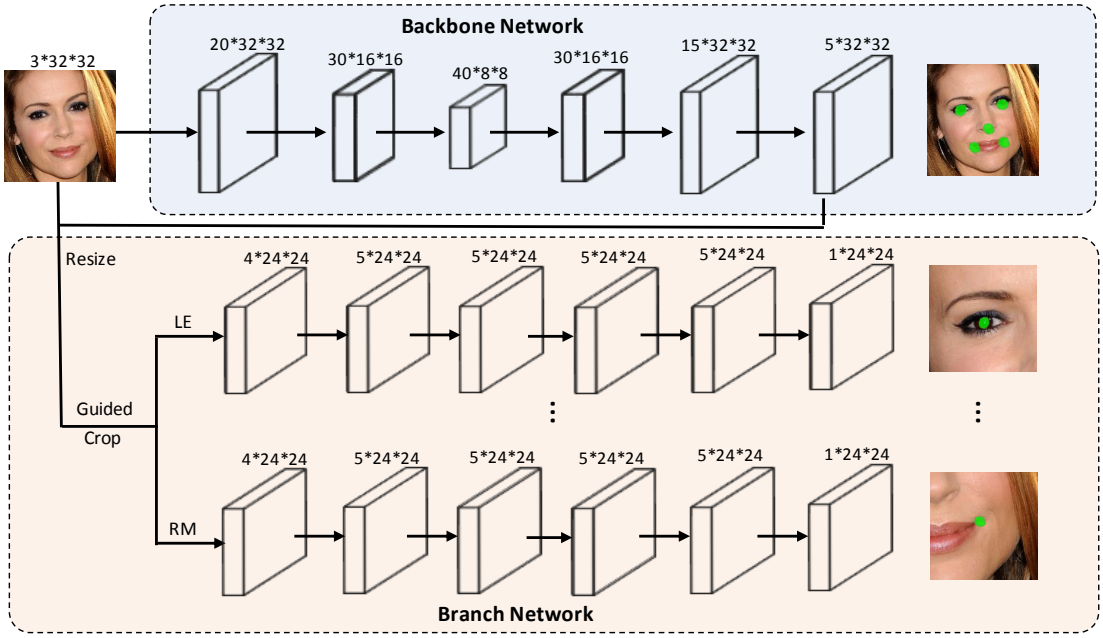

Fig. 2. The main architecture of the proposed Backbone-Branches Fully Convolutional Neural Network. It is capable of producing pixel-wise facial landmark response maps in a progressive way. The backbone network first generates low resolution response maps identifying rough landmark locations via a fully 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 landmarks and refines the related response map. Only down-sampling, up-sampling, and prediction layers are shown and intermediate convolutional layers are omitted in the network branches.

(LE), right eye (RE), nose (N), left mouth corner (LM) and right mouth corner (RM).

Unlike existing approaches that predict landmark locations by coordinate regression, we exploit fully convolutional neural networks (FCN) to directly produce response maps which indicate the probability of landmark existence at every image location. 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 while receptive fields without that type of landmarks 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 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 thus be defined as

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

where $I(P(x, y))$ denotes 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.

According to Equation (3), there is a trade-off between localization accuracy and computational cost. In order to achieve high accuracy, we need to compute response values for significantly overlapping receptive fields. However, in order to speed up the detection process, we should generate a coarser 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 [34] for image classification. More specifically, our 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.

In this paper, this two-component architecture is implemented as a backbone-branches fully convolutional neural network, 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,ch respectively represent the height, width and the number of channels of a kernel. We also use MP to denote a max-pooling layer. In our 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 is 2.

A. Backbone Network

The backbone network is a fully convolutional network. It efficiently generates an initial low-resolution response map for input image I. When localizing facial landmarks in an image taken in an 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:

$$\mathcal{L}_1(I;\mathbf{W}_c) = \sum_{k=1}^{K} ||H^k(I;\mathbf{W}_c) - H_c^k(I)||^2, \qquad (4)$$

where $H_c^k(I)$ is the groundtruth map for type-k landmarks.

To take into account the global context, such as geometric constraints among landmarks, the backbone network takes the entire image as the input and handles all facial landmarks together. During training, the input image patch is resized to 32×32. The backbone network is made up of eight convolutional layers 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 one responsible for detecting one type of landmarks. All the K branches are designed to share the same network structure. Take one branch as an example. Cropped patches of the original input image and regions from the backbone's output heat map are stacked together as its input. The input data therefore consists 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. In order to make the branch network better suited for landmark position refinement, we resize the original input image to 64×64, four times the size of the backbone's input, and at the same time zoom the heat map from the backbone network to 64×64 as well. 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 use $H(P;\mathbf{W}_f^k), H_0^k(P)$ to denote the heat map it generates and the corresponding groundtruth heat map of patch P, respectively. The loss function of this branch component is again defined as follows:

$$\mathcal{L}_2(P;\mathbf{W}_f^k) = ||H(P;\mathbf{W}_f^k) - H_0^k(P)||^2. \qquad (5)$$

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 kernel size equal to 5 and stride equal to 1 while the last convolutional layer consists of 5 channels with kernel size 1 and stride 1. As shown in Figure 2, each branch FCN component is detailed as follows: $C(5, 5\times5\times5)$ - $C(5, 5\times5\times5)$ - $C(5, 5\times5\times5)$ - $C(5, 5\times5\times5)$ - $C(1, 1\times1\times5)$.

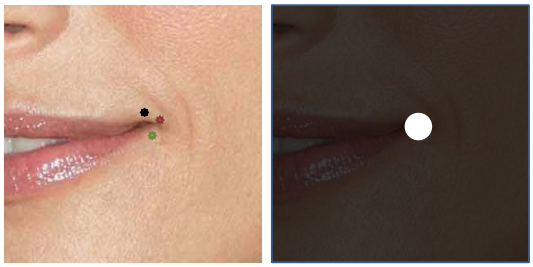

Fig. 3. (a) An isolated point cannot accurately reflect discrepancies among multiple annotations. The three points near the right mouth corner were annotated by three different workers. (b) We label a landmark as a small circular region rather than an isolated point in the groundtruth heat map.

C. Groundtruth Heat Map Generation

To our knowledge, the ground truth of a facial landmark is traditionally given as a single pixel location (x,y). To adapt such landmark specifications for the training stage of our proposed BB-FCN network, we generate the groundtruth 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 rest of the pixels. However, we argue that this method is suboptimal because an isolated point cannot reflect discrepancies among multiple annotations. As shown in Figure 3(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 an isolated point. We first initialize the heat map with zero everywhere, and then for each landmark p, we mark a circular region with center p and radius R in the groundtruth heat map with 1. Different radius is 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 as the backbone network estimates coarse landmark positions while the branch networks predict accurate landmark locations.

IV. Experimental Results

A. Datasets

To train our proposed BB-FCN, we collect 7317 face images (6317 for training, 1000 for validation) from the Internet and collect 7542 natural images (6542 for training, 1000 for validation) without any faces from Pascal-VOC2012 as negative samples. Each face is annotated with 72 landmarks. We use two public challenging datasets for evaluation: (AFW [10] and AFLW [11]. There is no overlap among the training, validation and evaluation datasets.

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. This dataset is intended for testing facial keypoint detection in unconstrained

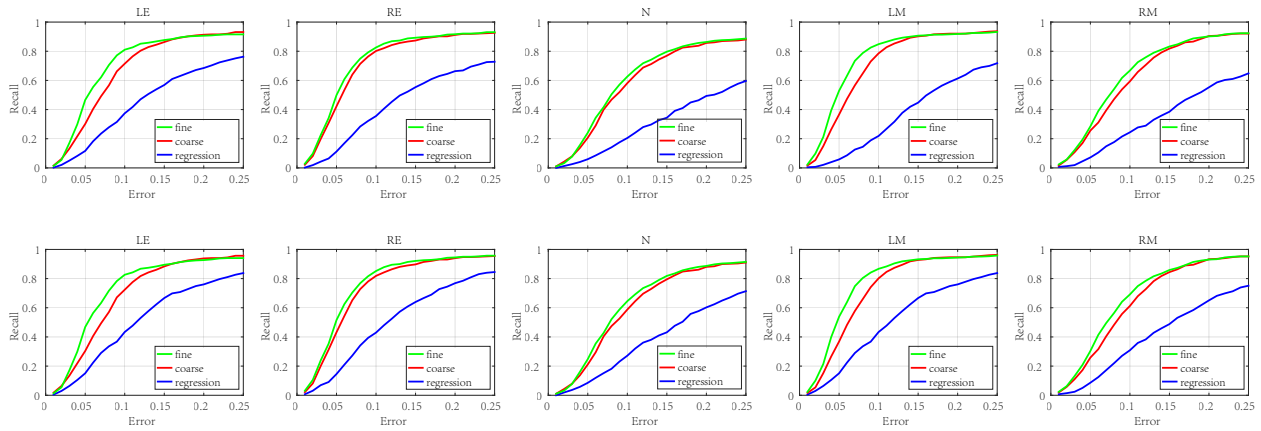

Fig. 4. The recall of landmarks on AFW in unconstrained settings. The curves labeled "fine" and "coarse" represent the performance of models with and without branch networks, respectively. The curve labeled "regression" represents the performance of the regression network based on a single fully convolutional 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.

settings, meaning faces may exhibit large pose, expression and illumination variations, and may have severe occlusions.

AFLW: This dataset contains 21,080 faces with large pose variations. It is very suitable for evaluating the performance of face alignment across a large range of poses. The selection of testing images from AFLW follows [25], which randomly chooses 3000 faces and 39% of them are non-frontal.

B. Implementation Details

We have implemented our proposed BB-FCN network in Caffe. During training, we initialize our networks by drawing weights from a zero-mean Gaussian distribution with a standard deviation equal to 0.01. The size of a mini-batch is 40. The positive training samples are image regions cropped from face images in our collected database. The Intersection-over-Union (IoU) between any cropped region and the original face image is above 0.5. The negative training samples are non-facial regions randomly cropped from the Pascal VOC 2012 dataset. Both the backbone and branch networks are trained using back-propagation and stochastic gradient descent (SGD) with 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 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 of landmark circles is set to 3% of the width of the input image.

During the testing phase, our BB-FCN network is able to accurately locate facial landmarks even without the assistance of a face detector. If a testing image is a cropped facial image, we resize it to match the size of training images and feed it to BB-FCN to produce the response heat map. When given an unconstrained image, we first construct an image pyramid and feed the images at different pyramid levels to the backbone network to generate multiple coarse heat maps. These heat maps are resized to match the size of the input image and fused to form a single heat map, which takes the maximum response across all original heat maps. Given a testing image, we build a pyramid of 20 levels by first resizing the image so that the length of the smaller side equal to 32 and gradually scaling it with 1.16 times every other layer for 20 times. The fused heat map is then used to generate candidate landmark regions fed into the branch networks. For each landmark type, we choose n locations with the highest response values from the output heat map of its branch network and take their average location as the final predicted landmark location, where n is the number of pixels in a landmark circle.

C. Evaluation Metric

To evaluate the accuracy of facial landmark localization, we adopt the mean (position) error as the metric. For a specific type of landmarks, 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 inter-ocular distance. The (position) error of a single landmark is defined as follows,

$$err = \frac{\sqrt{(x-x')^2+(y-y')^2}}{l} \times 100\%, \tag{6}$$

where (x, y) and (x', y') are the groundtruth and detected landmark locations, respectively, and the inter-ocular distance l is the Euclidean distance between the center points of the two eyes. In our experiments, we evaluate the mean error of every type of facial landmarks as well as 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) and A (average mean error of the five facial landmarks).

D. Performance Evaluation for Unconstrained Settings

Our BB-FCN is capable of dealing with facial images taken in unconstrained settings, e.g., the location of facial regions and the number of faces are unknown. We evaluate the performance of our BB-FCN using Recall-Error curves. A predictive facial landmark is considered correct if there exists a groundtruth landmark of the same type within the given position error. For a fixed number of predictive landmarks, the

Fig. 5. Qualitative facial landmark detection results in unconstrained settings. Our BB-FCN is capable of dealing with unconstrained facial images, even though the location of facial regions and the number of faces in the image are unknown. Best viewed in color.

recall rate (the fraction of ground truth annotations covered by predictive landmarks) varies as the acceptable position error increases, so that a Recall-Error curve can be obtained.

To our knowledge, very few facial landmark localization methods have been evaluated in the context of landmark detection in unconstrained settings. For the sake of fairness, we have also implemented a regression-based method using a fully convolutional network with nine convolutional layers. The setup of the first six layers is the same as in our backbone network while the denoted filters of the following two layers are $C(30, 2 \times 2 \times 40)$ and $C(30, 4 \times 4 \times 30)$ respectively. The top convolutional layer of the regression network produces a fifteen-channel output, with every three of which form a group. Each group of three channels indicates the probability of existence and the regressed two dimensional location of a landmark of a specific type. Given a threshold, this model can output the coordinates of all detected landmarks.

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 with one or both eyes are invisible, the inter-ocular distances are set up with 41.9% of the length of their annotated bounding boxes. [1] Figure 4 shows the Recall-Error curves of different types of landmarks, where the curves labeled "fine" and "coarse" represent the performance of our complete BB-FCN model and the backbone network alone, respectively. The curve labeled "regression" represents the performance of the above regression network based on a single FCN. Our methods significantly outperform the regression network, and the complete BB-FCN model performs much better than the backbone network alone. With a prediction of 15 landmarks for each landmark type, the complete model recalls 45% more landmarks than the regression network, when the acceptable position error is set within 8% of the inter-ocular distance. As the number of landmark prediction of each type increases to 30, the recalls of five landmarks within a position error of 25% of the inter-ocular distance are 94.1%, 95.7%, 91.5%, 95.8% and 95.2% respectively. Given more predicted landmarks, we can achieves higher landmarks recalls. Figure 5 demonstrates some landmark detection results on the AFW dataset in unconstrained settings.

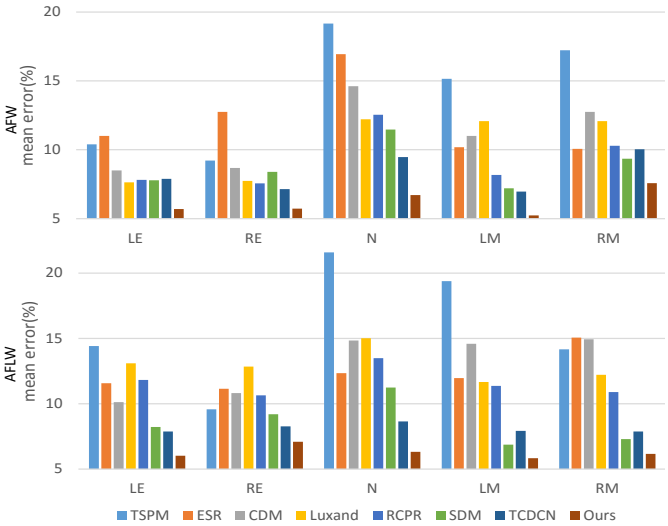

Fig. 6. Comparisons with state-of-the-art methods on two public datasets. The top row shows corresponding results on AFW, and the bottom row shows corresponding results on AFLW. The average mean errors of all participating methods are summarized in Table I.

E. Comparison with the State of the Art

We compare our method with other state-of-the-art methods, i.e., (1) Robust Cascaded Pose Regression (RCPR) [35]; (2) Tree Structured Part Model (TSPM) [10]; (3) Luxand face SDK [2]; (4) Explicit Shape Regression (ESR) [20]; (5) A Cascaded Deformable Shape Model (CDM) [36]; (6) Supervised

[1]The average ratio between the inter-ocular distances of the common faces and the length of their annotated bounding boxes is 41.9% on AFW.

[2]Luxand face SDK: http://www.luxand.com/

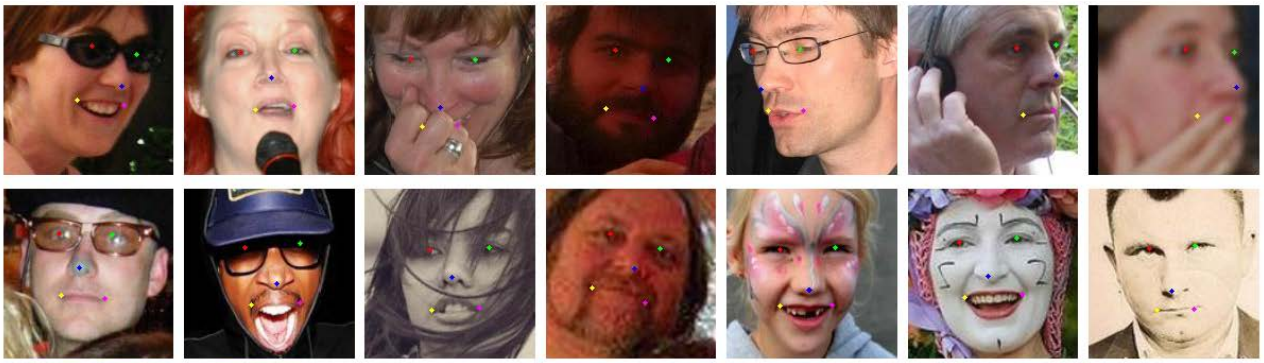

Fig. 7. Qualitative facial landmark localization results by our method. The first row shows the results on AFW and the second row shows the results on AFLW. Our method is robust under occlusion, exaggerated expressions and extreme illumination.

TABLE I
AVERAGE MEAN ERRORS OF OUR METHOD AND 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
Ours	6.18	6.28

TABLE II
AVERAGE MEAN ERRORS OF THE COMPLETE BACKBONE-BRANCHES NETWORK AND THE BACKBONE NETWORK ALONE 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

Descent Method (SDM) [37]; (7) Tasks-Constrained Deep Convolutional Network (TCDCN) [25]; (8) Multi-task Cascaded Convolutional Networks (MTCNN) [38]; (9) Recurrent Attentive-Refinement Networks (RAR) [29]. The results of some competing methods are quoted from [25].

On the AFW dataset, our average mean error over five landmark types is 6.18%, which improves 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, 21.5% improvement over TDCN. Table I demonstrates that our BB-FCN network outperforms all competing methods on the three datasets. Qualitative results in Figure 7 show that our method is robust under occlusion, exaggerated expressions and extreme illumination.

F. Ablation Study

Our proposed BB-FCN is composed of two components, the backbone network and the branch networks. To show

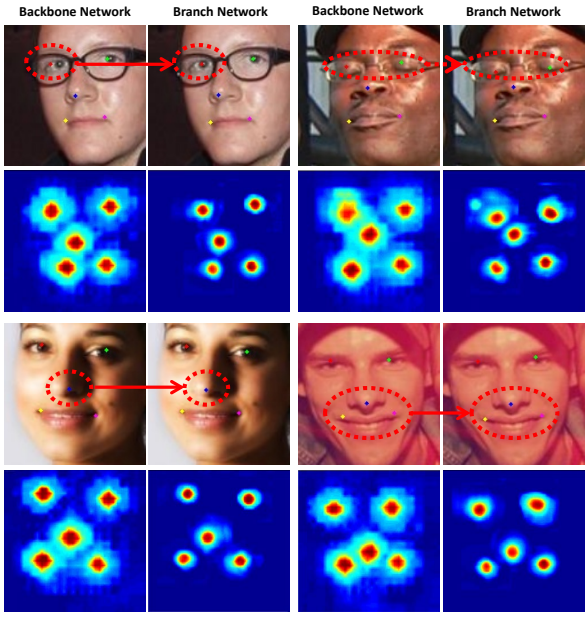

Fig. 8. Examples of improvements made by the branch networks. The response heat maps of the branch networks are more compact and precise. Best viewed in color.

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 II, the average mean error on AFLW is decreased from 8.31% to 6.28%, with about 24.4% relative improvement, after the branch networks are added to perform landmark refinement. Figure 8 demonstrates visual improvements achieved with the branch networks over the single backbone network. We can see that the output heat maps of the branch networks are more compact and precise than those of the backbone network.

G. Runtime Efficiency

One of the most important characteristics of landmark and face detectors is their runtime efficiency. Our method performs accurate and efficient detection via a coarse-to-fine pipeline. Table III shows the running time of several deep models for facial landmark detection. Among these models, TCDCN requires 18ms to process a facial image on an Intel Core i5

TABLE III
COMPARISON OF RUNNING TIMES ON CPU AMONG DEEP MODELS FOR FACIAL LANDMARK DETECTION.

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

CPU, which is 7 times faster than CDCN [39]. CFAN [27] needs 30ms to run multiple auto-encoders. Our method only needs 9ms on an Intel Core i5 2.80GHz CPU and 1.8ms on a NVIDIA Titan X GPU. Our method also achieves practical runtime efficiency under unconstrained settings. To locate facial landmarks not smaller than 80×80 in 640×480 VGA images, our landmark detector can run at 30 FPS.

V. CONCLUSIONS

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. Specifically, our architecture contains a backbone network for roughly detecting the locations of all facial landmarks and one branch network for each type of detected landmarks for further refining their locations. Our extensive experiments demonstrate that BB-FCN achieves very promising results on both traditional benchmarks with a controlled setting as well as cluttered, real-world scenes.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant 61702565, Guangdong Natural Science Foundation Project for Research Teams under Grant 2017A030312006, and was also sponsored by CCF-Tencent Open Research Fund.

REFERENCES

[1] C. Lu and X. Tang, "Surpassing human-level face verification performance on lfw with gaussianface," in *AAAI*, 2015.
[2] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep learning identity-preserving face space," in *ICCV*, 2013, pp. 113–120.
[3] Y. Li, L. Liu, L. Lin, and Q. Wang, "Face recognition by coarse-to-fine landmark regression with application to atm surveillance," in *CCCV*, 2017.
[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014, pp. 580–587.
[5] G. Li and Y. Yu, "Contrast-oriented deep neural networks for salient object detection," *IEEE Transactions on Neural Networks and Learning Systems*, 2018.
[6] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li, "Disc: Deep image saliency computing via progressive representation learning." *TNNLS*, vol. 27, no. 6, pp. 1135–1149, 2016.
[7] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," 2017.
[8] L. Liu, H. Wang, G. Li, W. Ouyang, and L. Lin, "Crowd counting using deep recurrent spatial-aware network," in *IJCAI*, 2018.
[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.
[10] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *CVPR*. IEEE, 2012, pp. 2879–2886.
[11] 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 *ICCV Workshops*. IEEE, 2011, pp. 2144–2151.
[12] P. Sauer, T. F. Cootes, and C. J. Taylor, "Accurate regression procedures for active appearance models." in *BMVC*, 2011, pp. 1–11.
[13] P. A. Tresadern, P. Sauer, and T. F. Cootes, "Additive update predictors in active appearance models." in *BMVC*, vol. 2. Citeseer, 2010, p. 4.
[14] L. Liang, R. Xiao, F. Wen, and J. Sun, "Face alignment via component-based discriminative search," in *ECCV*. Springer, 2008, pp. 72–85.
[15] G. Tzimiropoulos and M. Pantic, "Optimization problems for fast aam fitting in-the-wild," in *ICCV*, 2013, pp. 593–600.
[16] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *PAMI*, no. 6, pp. 681–685, 2001.
[17] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *IJCV*, vol. 91, no. 2, pp. 200–215, 2011.
[18] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *PAMI*, vol. 35, no. 12, pp. 2930–2940, 2013.
[19] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool, "Real-time facial feature detection using conditional regression forests," in *CVPR*. IEEE, 2012, pp. 2578–2585.
[20] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *IJCV*, vol. 107, no. 2, pp. 177–190, 2014.
[21] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *CVPR*, 2014, pp. 1685–1692.
[22] S. Zhu, C. Li, C.-C. Loy, and X. Tang, "Unconstrained face alignment via cascaded compositional learning," in *CVPR*, 2016, pp. 3409–3417.
[23] O. Tuzel, T. K. Marks, and S. Tambe, "Robust face alignment using a mixture of invariant experts," in *ECCV*. Springer, 2016, pp. 825–841.
[24] X. Fan, R. Liu, Z. Luo, Y. Li, and Y. Feng, "Explicit shape regression with characteristic number for facial landmark localization," *TMM*, 2017.
[25] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *ECCV*. Springer, 2014, pp. 94–108.
[26] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade," in *ICCV Workshops*, 2013, pp. 386–391.
[27] J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment," in *ECCV*. Springer, 2014, pp. 1–16.
[28] X. Peng, R. S. Feris, X. Wang, and D. N. Metaxas, "A recurrent encoder-decoder network for sequential face alignment," in *ECCV*. Springer, 2016, pp. 38–56.
[29] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim, "Robust facial landmark detection via recurrent attentive-refinement networks," in *ECCV*. Springer, 2016, pp. 57–72.
[30] 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 *CVPR*, 2016, pp. 4177–4187.
[31] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *CVPR*, 2016, pp. 146–155.
[32] A. Jourabloo and X. Liu, "Large-pose face alignment via cnn-based dense 3d model fitting," in *CVPR*, 2016, pp. 4188–4196.
[33] F. Liu, D. Zeng, Q. Zhao, and X. Liu, "Joint face alignment and 3d face reconstruction," in *ECCV*. Springer, 2016, pp. 545–560.
[34] Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, and Y. Yu, "Hd-cnn: Hierarchical deep convolutional neural networks for large scale visual recognition," in *ICCV*, 2015, pp. 2740–2748.
[35] X. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *ICCV*, 2013, pp. 1513–1520.
[36] 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 *ICCV*, 2013, pp. 1944–1951.
[37] X. Xiong and F. Torre, "Supervised descent method and its applications to face alignment," in *CVPR*, 2013, pp. 532–539.
[38] 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.
[39] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *CVPR*, 2013, pp. 3476–3483.