From Facial Parts Responses to Face Detection: A Deep Learning Approach

Shuo Yang[1,2] Ping Luo[2,1] Chen Change Loy[1,2] Xiaoou Tang[1,2]
[1]Department of Information Engineering, The Chinese University of Hong Kong
[2]Shenzhen Key Lab of Comp. Vis. & Pat. Rec., Shenzhen Institutes of Advanced Technology, CAS, China
{ys014, pluo, ccloy, xtang}@ie.cuhk,edu.hk

Abstract

In this paper, we propose a novel deep convolutional network (DCN) that achieves outstanding performance on FDDB, PASCAL Face, and AFW. Specifically, our method achieves a high recall rate of 90.99% on the challenging FDDB benchmark, outperforming the state-of-the-art method [23] by a large margin of 2.91%. Importantly, we consider finding faces from a new perspective through scoring facial parts responses by their spatial structure and arrangement. The scoring mechanism is carefully formulated considering challenging cases where faces are only partially visible. This consideration allows our network to detect faces under severe occlusion and unconstrained pose variation, which are the main difficulty and bottleneck of most existing face detection approaches. We show that despite the use of DCN, our network can achieve practical runtime speed.

1. Introduction

Neural network based methods were once widely applied for localizing faces [33, 26, 7, 25], but they were soon replaced by various non-neural network-based face detectors, which are based on cascade structure [3, 9, 20, 34] and deformable part models (DPM) [23, 36, 40] detectors. Deep convolutional networks (DCN) have recently achieved remarkable performance in many computer vision tasks, such as object detection, object classification, and face recognition. Given the recent advances of deep learning and graphical processing units (GPUs), it is worthwhile to revisit the face detection problem from the neural network perspective.

In this study, we wish to design a deep convolutional network for face detection, with the aim of not only exploiting the representation learning capacity of DCN, but also formulating a novel way for handling the severe occlusion issue, which has been a bottleneck in face detection. To this end, we design a new deep convolutional network with the following appealing properties: (1) It is robust to severe occlusion. As depicted in Fig. 1, our method can detect faces even more than half of the face region is occluded; (2) it is capable of detecting faces with large pose variation, *e.g.* profile view without training separate models under different viewpoints; (3) it accepts full image of arbitrary size and the faces of different scales can appear anywhere in the image.

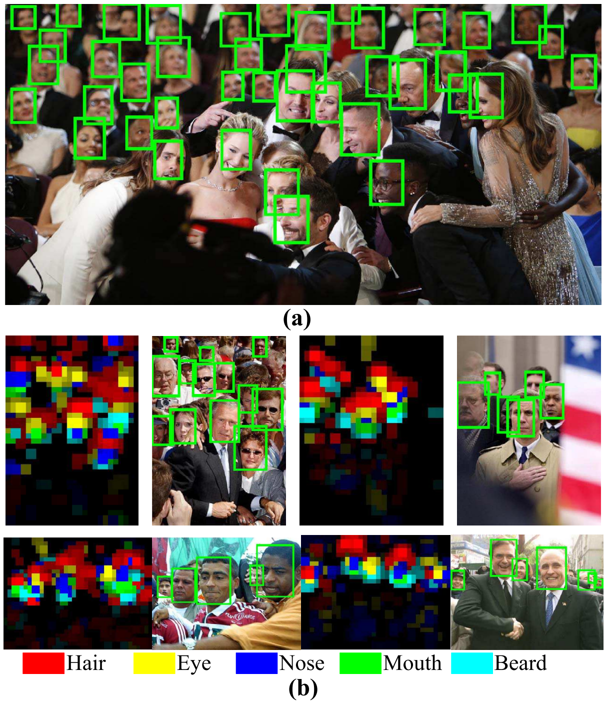


Figure 1. (a) We propose a deep convolutional network for face detection, which achieves high recall of faces even under severe occlusions and head pose variations. The key to the success of our approach is the new mechanism for scoring face likeliness based on deep network responses on local facial parts. (b) The part-level response maps (we call it 'partness' map) generated by our deep network given a full image without prior face detection. All these occluded faces are difficult to handle by conventional approach.

All the aforementioned properties, which are challenging to achieve with conventional approaches, are made possible with the following considerations:

(1) *Generating face parts responses from attribute-aware deep networks*: We believe the reasoning of unique structure of local facial parts (*e.g*. eyes, nose, mouths) is the key to address face detection in unconstrained environment. To this end, we design a set of attribute-aware deep networks, which are pre-trained with generic objects and then fine-tuned with specific part-level binary attributes (*e.g*. mouth attributes including big lips, opened mouth, smiling, wearing lipstick). We show that these networks could generate response maps in deep layers that strongly indicate the locations of the parts. The examples depicted in Fig. 1(b) show the responses maps (known as 'partness map' in our paper) of five different face parts.

(2) *Computing faceness score from responses configurations*: Given the parts responses, we formulate an effective method to reason the degree of face likeliness through analysing their spatial arrangement. For instance, the hair should appear above the eyes, and the mouth should only appear below the nose. Any inconsistency would be penalized. Faceness scores will be derived and used to re-rank candidate windows of any generic object proposal generator to obtain a set of face proposals. Our experiment shows that our face proposal enjoys a high recall with just modest number of proposals (over 90% of face recall with around 150 proposals, ≈0.5% of full sliding windows, and ≈10% of generic object proposals).

(3) *Refining the face hypotheses* – Both the aforementioned components offer us the chance to find a face even under severe occlusion and pose variations. The output of these components is a small set of high-quality face bounding box proposals that cover most faces in an image. Given the face proposals, we design a multitask deep convolutional network in the second stage to refine the hypotheses further, by simultaneously recognizing the true faces and estimating more precise face locations.

Our main contribution in this study is the novel use of DCN for discovering facial parts responses from arbitrary uncropped face images. Interestingly, in our method, part detectors emerge within CNN trained to classify attributes from uncropped face images, without any part supervision. This is new in the literature. We leverage this new capability to further propose a face detector that is robust to severe occlusion. Our network achieves the state-of-the-art performance on challenging face detection benchmarks including FDDB, PASCAL Faces, and AFW. We show that practical runtime speed can be achieved albeit the use of DCN.

2. Related Work

There is a long history of using neural network for the task of face detection [33, 26, 7, 25]. An early face detection survey [38] provides an extensive coverage on relevant methods. Here we highlight a few notable studies. Rowley *et al*. [26] exploit a set of neural network-based filters to detect presence of faces in multiple scales, and merge the detections from individual filters. Osadchy *et al*. [25] demonstrate that a joint learning of face detection and pose estimation significantly improves the performance of face detection. The seminal work of Vaillant *et al*. [33] adopt a two-stage coarse-to-fine detection. Specifically, the first stage approximately locates the face region, whilst the second stage provides a more precise localization. Our approach is inspired by these studies, but we introduce innovations on many aspects. In particular, we employ contemporary deep learning strategies, *e.g*. pre-training, to train deeper networks for more robust feature representation learning. Importantly, our first stage network is conceptually different from that of [33], and many recent deep learning detection frameworks – we train attribute-aware deep convolutional networks to achieve precise localization of facial parts, and exploit their spatial structure for inferring face likeliness. This concept is new and it allows our model to detect faces under severe occlusion and pose variations. While great efforts have been devoted for addressing face detection under occlusion [21, 22], these methods are all confined to frontal faces. In contrast, our model can discover faces under variations of both pose and occlusion.

In the last decades, cascade based [3, 9, 20, 34] and deformable part models (DPM) detectors dominate the face detection approaches. Viola and Jones [34] introduced fast Haar-like features computation via integral image and boosted cascade classifier. Various studies thereafter follow a similar pipeline. Amongst the variants, SURF cascade [20] was one of the top performers. Later Chen *et al*. [3] demonstrate state-of-the-art face detection performance by learning face detection and face alignment jointly in the same cascade framework. Deformable part models define face as a collection of parts. Latent Support Vector Machine is typically used to find the parts and their relationships. DPM is shown more robust to occlusion than the cascade based methods. A recent study [23] demonstrates state-of-the-art performance with just a vanilla DPM, achieving better results than more sophisticated DPM variants [36, 40].

A recent study [6] shows that face detection can be further improved by using deep learning, leveraging the high capacity of deep convolutional networks. In this study, we push the performance limit further. Specifically, the network proposed by [6] does not have explicit mechanism to handle occlusion, the face detector therefore fails to detect faces with heavy occlusions, as acknowledged by the authors. In contrast, our two-stage architecture has its first stage designated to handle partial occlusions. In addition, our network gains improved efficiency by adopting the more recent fully convolutional architecture, in contrast to the previous work that relies on the conventional sliding window approach to obtain the final face detector.

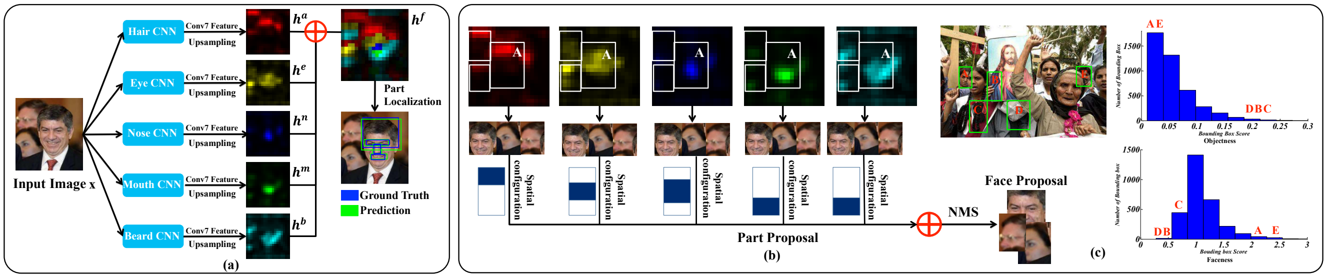


Figure 2. (a) The pipeline of generating part response maps and part localization. Different CNNs are trained to handle different facial parts, but they can share deep layers for computational efficiency. (b) The pipeline for generating face proposals. (c) Bounding box reranking by face measure **(Best viewed in color).**

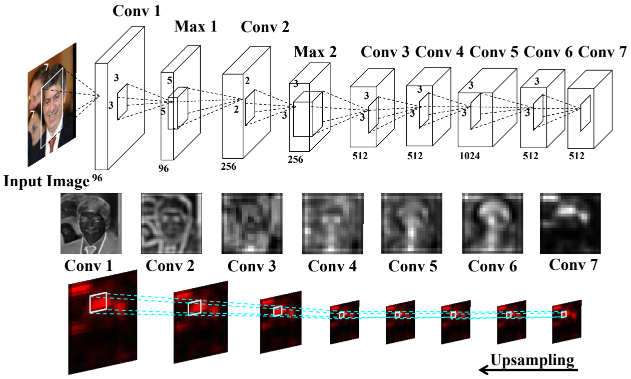


Figure 3. A general architecture of an attribute-aware deep network. Other architecture is possible.

The first stage of our model is partially inspired by the generic object proposal approaches [2, 32, 41]. Generic object proposal generators are now an indispensable component of standard object detection algorithms through providing high-quality and category-independent bounding boxes. These generic methods, however, are devoted to generic objects therefore not suitable to propose windows specific to face. In particular, applying a generic proposal generator directly would produce enormous number of candidate windows but only minority of them contain faces. In addition, a generic method does not consider the unique structure and parts on the face. Hence, there will be no principled mechanism to recall faces when the face is only partially visible. These shortcomings motivate us to formulate the new faceness measure to achieve high recall on faces, whilst reduce the number of candidate windows to half the original.

3. Faceness-Net

This section introduces the proposed attribute-aware face proposal and face detection approach, *Faceness-Net*. In the following, we first briefly overview the entire pipeline and then discuss the details.

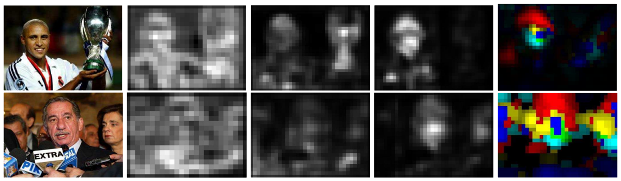
Faceness-Net's pipeline consists of three stages, *i.e.* generating partness maps, ranking candidate windows by faceness scores, and refining face proposals for face detection. In the first stage as shown in Fig. 2(a), a full image $\mathbf{x}$ is used as input to five CNNs. Note that all the five CNNs can share deep layers to save computational time. Each CNN outputs a partness map, which is obtained by weighted averaging over all the label maps at its top convolutional layer. Each of these partness maps indicates the location of a specific facial component presented in the image, *e.g.* hair, eyes, nose, mouth, and beard, denoted by $\mathbf{h}^a$, $\mathbf{h}^e$, $\mathbf{h}^n$, $\mathbf{h}^m$, and $\mathbf{h}^b$, respectively. We combine all these partness maps into a face label map $\mathbf{h}^f$, which clearly designates faces' locations.

In the second stage, given a set of candidate windows that are generated by existing object proposal methods such as [2, 32, 41], we rank these windows according to their faceness scores, which are extracted from the partness maps with respect to different facial parts configurations, as illustrated at the bottom of Fig. 2(b). For example, as visualized in Fig. 2(b), a candidate window 'A' covers a local region of $\mathbf{h}^a$ (*i.e.* hair) and its faceness score is calculated by dividing the values at its upper part with respect to the values at its lower part, because hair is more likely to present at the top of a face region. A final faceness score of 'A' is obtained by averaging over the scores of these parts. In this case, large number of false positive windows can be pruned. Notably, the proposed approach is capable of coping with severe face occlusions, as shown in Fig. 2(c), where face windows 'A' and 'E' can be retrieved by objectness [1] only if large amount of windows are proposed, whilst they rank top 50 by using our method.

In the last stage, the proposed candidate windows are refined by training a multitask CNN, where face classification and bounding box regression are jointly optimized.

3.1. Partness Maps Extraction

Network structure. Fig. 3 depicts the structure and hyper-parameters of the CNN in Fig. 2(a), which stacks seven convolutional layers (conv1 to conv7) and two max-pooling layers (max1 and max2). This convolutional structure is inspired by the AlexNet [16] in image classification. Many recent studies [29, 39] showed that stacking many convolutions as AlexNet did can roughly capture object locations.



(a) Original image (b) Without pre-trained model (c) Fine-tuned by classifying face vs. non-face images (d) Fine-tuned by classifying 25 face attributes (e) Fine-tuned by classifying grouped face attributes w.r.t facial parts

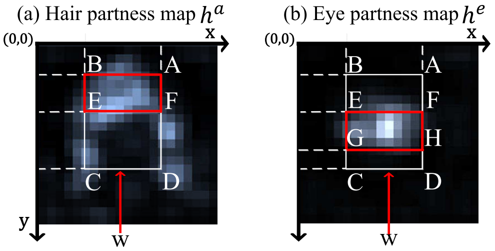
Figure 4. The responses or partness maps obtained by using different types of supervisions.

Table 1. Facial attributes grouping.

Facial Part	Facial Attributes
Hair	Black hair, Blond hair, Brown hair, Gray hair, Bald, Wavy hair, Straight hair, Receding hairline, Bangs
Eye	Bushy eyebrows, Arched eyebrows, Narrow eyes, Bags under eyes, Eyeglasses
Nose	Big nose, Pointy nose
Mouth	Big lips, Mouth slightly open, Smiling, Wearing lipstick
Beard	No beard, Goatee, 5 o'clock shadow, Mustache, Sideburns

Learning partness maps. As shown in Fig. 4(b), a deep network trained on generic objects, *e.g*. AlexNet [16], is not capable of providing us with precise faces' locations, let alone partness map. The partness maps can be learned in multiple ways. The most straight-forward manner is to use the image and its pixelwise segmentation label map as input and target, respectively. This setting is widely employed in image labeling [5, 24]. However, it requires label maps with pixelwise annotations, which are expensive to collect. Another setting is image-level classification (*i.e*. faces and non-faces), as shown in Fig. 4(c). It works well where the training images are well-aligned, such as face recognition [30]. Nevertheless, it suffers from complex background clutter because the supervisory information is not sufficient to account for face variations. Its learned feature maps contain too much noises, which overwhelm the actual faces' locations. Attribute learning in Fig. 4(d) extends the binary classification in (c) to the extreme by using a combination of attributes to capture face variations. For instance, an 'Asian' face can be distinguished from a 'European' face. However, our experiments demonstrate that the setting is not robust to occlusion. Hence, as shown in Fig. 4(e), this work extends (d) by partitioning attributes into groups based on facial components. For instance, 'black hair', 'blond hair', 'bald', and 'bangs' are grouped together, as all of them are related to hair. The grouped attributes are summarized in Table 1. In this case, different face parts can be modeled by different CNNs (with option to share some deep layers). If one part is occluded, the face region can still be localized by CNNs of the other parts.

We take the Hair-CNN in Fig. 2(a) as an example to illustrate the learning procedure. Let $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$ be a set of full face images and the attribute labels of hair, where $\forall \mathbf{x}_i \in \mathbb{R}^{256\times256}$ and $\forall \mathbf{y}_i \in \mathbb{R}^{1\times9}$, implying that each full image is rescaled to 256 × 256 and there is nine attributes related to hair as listed in Table 1. Learning is formulated as a multi-variate classification problem by minimizing the cross-entropy loss, $L = \sum_{i=1}^N \mathbf{y}_i \log p(\mathbf{y}_i = 1|\mathbf{x}_i) + (\mathbf{1} - \mathbf{y}_i) \log\big(\mathbf{1} - p(\mathbf{y}_i = 1|\mathbf{x}_i)\big)$, where $p(\mathbf{y}_i|\mathbf{x}_i)$ is modeled as a sigmoid function, indicating the probability of the presence of the attributes. This loss function can be optimized by the stochastic gradient descent with back-propagation.


(a) Hair partness map h^a

(b) Eye partness map h^e

Figure 5. Examples of spatial configurations (**Best viewed in color**).

However, the partness map generated by the Hair-CNN trained as above contains erroneous responses at the background, revealing that this training scheme is not sufficient to account for background clutter. To obtain a cleaner partness map, we employ the merit from object categorization, where CNN is pre-trained with massive general object categories in ImageNet [27] as in [16]. It can be viewed as a supervised pre-training for the Hair-CNN.

3.2. Ranking Windows by Faceness Measure

Our approach is loosely coupled with existing generic object proposal generators [2, 32, 41] - it accepts candidate windows from the latter but generates its own faceness measure to return a ranked set of top-scoring face proposals. Fig. 5 takes hair and eyes to illustrate the procedure of deriving the faceness measure from a partness map. Let Δ_w be the faceness score of a window w. For example, as shown in Fig. 5(a), given a partness map of hair, $\mathbf{h}^a$, Δ_w is attained by dividing the sum of values in ABEF (red) by the sum of values in FECD. Similarly, Fig. 5(b) expresses that Δ_w is obtained by dividing the sum of values in EFGH (red) with respect to ABEF+HGCD of $\mathbf{h}^e$.

For both of the above examples, larger value of Δ_w indicates w has higher overlapping ratio with face. These spatial configurations, such as ABEF in (a) and EFGH in (b), can be learned from data. We take hair as an example. We need to learn the positions of points E and F, which can be represented by the (x, y)-coordinates of ABCD, *i.e*. the proposed window. For instance, the position of E in (a) can be represented by $x_e = x_b$ and $y_e = \lambda y_b + (1 - \lambda) y_c$, implying that the value of its y-axis is a linear combination of y_b and y_c.

With this representation, Δ_w can be efficiently computed by using the integral image (denoted as $\mathbf{I}$) of the partness map. For instance, Δ_w in (a) is attained by

$$\frac{\mathbf{I}(x_f,y_f)+\mathbf{I}(x_b,y_b)-\mathbf{I}(x_a,y_a)-\mathbf{I}\big(x_b,\lambda y_b+(1-\lambda)y_c\big)}{\mathbf{I}(x_d,y_d)+\mathbf{I}(x_e,y_e)-\mathbf{I}\big(x_a,\lambda y_a+(1-\lambda)y_d\big)-\mathbf{I}(x_c,y_c)}, \tag{1}$$

where $\mathbf{I}(x,y)$ signifies the value at the location (x,y).

Given a training set $\{w_i, r_i, \mathbf{h}_i\}_{i=1}^{M}$, where w_i and $r_i \in \{0,1\}$ denote the i-th window and its label (*i.e.* face/non-face), respectively. $\mathbf{h}_i$ is the cropped partness map with respect to the i-th window, *e.g.* region ABCD in $\mathbf{h}^a$. This problem can be simply formulated as maximum a posteriori (MAP)

$$\lambda^* = \arg\max_{\lambda} \prod_{i}^{M} p(r_i|\lambda, w_i, \mathbf{h}_i)p(\lambda, w_i, \mathbf{h}_i), \tag{2}$$

where λ represents a set of parameters when learning the spatial configuration of hair (Fig. 5(a)). $p(r_i|\lambda, w_i, \mathbf{h}_i)$ and $p(\lambda, w_i, \mathbf{h}_i)$ stand for the likelihood and prior, respectively. The likelihood of faceness can be modeled by a sigmoid function, *i.e.* $p(r_i|\lambda, w_i, \mathbf{h}_i) = \frac{1}{1+\exp(\frac{-\alpha}{\Delta_{w_i}})}$, where α is a coefficient. This likelihood measures the confidence of partitioning face and non-face, given a certain spatial configuration. The prior term can be factorized, $p(\lambda, w_i, \mathbf{h}_i) = p(\lambda)p(w_i)p(\mathbf{h}_i)$, where $p(\lambda)$ is a uniform distribution between zero and one, as it indicates the coefficients of linear combination, $p(w_i)$ models the prior of the candidate window, which can be generated by object proposal methods, and $p(\mathbf{h}_i)$ is the partness map as in Sec. 3.1. Since λ typically has low dimension in this work (*e.g.* one dimension of hair), it can be simply obtained by line search. Nevertheless, Eq.(2) can be easily extended to model more complex spatial configurations.

3.3. Face Detection

The proposed windows achieved by faceness measure have high recall rate. To improve it further, we refine these windows by joint training face classification and bounding box regression using a CNN similar to the AlexNet [16].

In particular, we fine-tune AlexNet using face images from AFLW [15] and person-free images from PASCAL VOC 2007 [4]. For face classification, a proposed window is assigned with a positive label if the IoU between it and the ground truth bounding box is larger than 0.5; otherwise it is negative. For bounding box regression, each proposal is trained to predict the positions of its nearest ground truth bounding box. If the proposed window is a false positive, the CNN outputs a vector of $[-1,-1,-1,-1]$. We adopt the Euclidean loss and cross-entropy loss for bounding box regression and face classification, respectively.

4. Experimental Settings

Training datasets. (i) We employ CelebFaces dataset [31] to train our attribute-aware networks. The dataset contains 87,628 web-based images exclusive from the LFW [10], FDDB [12], AFW [40] and PASCAL [36] datasets. We label all images in the CelebFaces dataset with 25 facial attributes and divide the labeled attributes into five categories based on their respective facial parts as shown in Table 1. We randomly select $75,000$ images from the CelebFaces dataset for training and the remaining is reserved as validation set. (ii) For face detection training, we choose $13,205$ images from the AFLW dataset [15] to ensure a balanced out-of-plane pose distribution and $5,771$ random person-free images from the PASCAL VOC 2007 dataset.

Part response testing dataset. In Sec. 5.1, we use LFW dataset [10] for evaluating the quality of part response maps for part localization. We select $2,927$ LFW images following [14] since it provides manually labeled hair+beard superpixel labels, on which the minimal and maximal coordinates can be used to generate the ground truth of face parts bounding boxes. Similarly, face parts boxes for eye, nose and mouth are manually labeled guided by the 68 dense facial landmarks.

Face proposal and detection testing datasets. In Sec. 5.2 and Sec. 5.3, we use the following datasets. (i) FDDB [12] dataset contains the annotations for $5,171$ faces in a set of $2,845$ images. For the face proposal evaluation, we follow the standard evaluation protocol in object proposal studies [41] and transform the original FDDB ellipses ground truth into bounding boxes by minimal bounding rectangle. For the face detection evaluation, the original FDDB ellipse ground truth is used. (ii) AFW [40] dataset is built using Flickr images. It has 205 images with 473 annotated faces with large variations in both face viewpoint and appearance. (iii) PASCAL faces [36] is a widely used face detection benchmark dataset. It consists of 851 images and $1,341$ annotated faces.

Evaluation settings. Following [41], we employ the Intersection over Union (IoU) as evaluation metric. We fix the IoU threshold to 0.5 following the strict PASCAL criterion. In particular, an object is considered being covered/detected by a proposal if IoU is no less than 0.5. To evaluate the effectiveness of different object proposal algorithms, we use the detection rate (DR) given the number of proposals per image [41]. For face detection, we use standard precision and recall (PR) to evaluate the effectiveness of face detection algorithms.

5. Results

5.1. Evaluating the Quality of Partness Maps

Robustness to unconstrained training input. In the testing stage, the proposed approach does not assume well-

Table 2. Facial part detection rate. The number of proposals = 350.

Training Data	Hair	Eye	Nose	Mouth
Cropped	95.56%	95.87%	92.09%	94.17%
Uncropped	94.57%	97.19%	91.25%	93.55%

cropped faces as input. In the training stage, our approach neither requires well-cropped faces for learning. This is an unique advantage over existing approaches.

To support this statement, we conduct an experiment by fine-tuning two different CNNs as in Fig. 2(a), each of which taking different inputs: (1) uncropped images, which may include large portion of background clutters apart the face; and (2) cropped images, which encompass roughly the face and shoulder regions. The performance is measured based on the part detection rate[1]. Note that we combine the evaluation on 'Hair+Beard' to suit the ground truth provided by [14] (see Sec. 4). The detection results are summarized in Table 2. As can be observed, the proposed approach performs similarly given both the uncropped and cropped images as training inputs. The results suggest the robustness of the method in handling unconstrained images for training. In particular, thanks to the facial attribute-driven training, despite the use of uncropped images, the deep model is encouraged to discover and capture the facial part representation in the deep layers, it is therefore capable of generating response maps that precisely pinpoint the locations of parts. In the following experiments, all the proposed models are trained on uncropped images. Fig. 6(a) shows the qualitative results. Note that facial parts can be discovered despite challenging poses.

5.2. From Part Responses to Face Proposal

Comparison with generic object proposals. In this experiment, we show the effectiveness of adapting different generic object proposal generators [2, 41, 32] to produce face-specific proposals. Since the notion of face proposal is new, no suitable methods are comparable therefore we use the original generic methods as baselines. We first apply any object proposal generator to generate the proposals and we use our method described in Sec. 3.2 to obtain the face proposals. We experiment with different parameters for the generic methods, and choose parameters that produce moderate number of proposals with very high recall. Evaluation is conducted following the standard protocol [41].

The results are shown in Fig. 7. It can be observed that our method consistently improves the state-of-the-art methods for proposing face candidate windows, under different IoU thresholds. Table 3 shows that our method achieves high recall with small number of proposals.

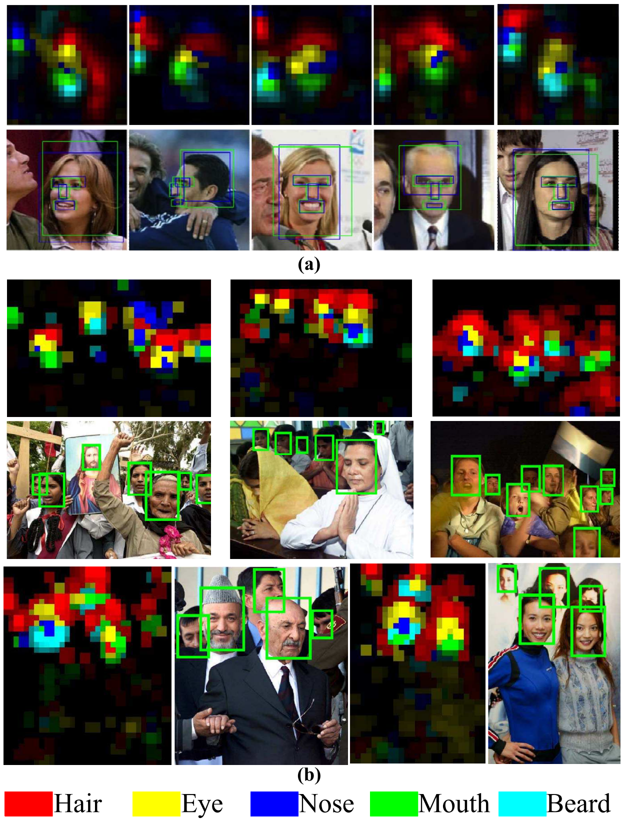


Figure 6. (a) The top row depicts the response maps generated by the proposed approach on each part. The second row shows the part localization results. Ground truth is depicted by the blue bounding boxes, whilst our part proposals are indicated in green. (b) Face detection results on FDDB images. The bounding box in green is detected by our method. We show the partness maps as reference.

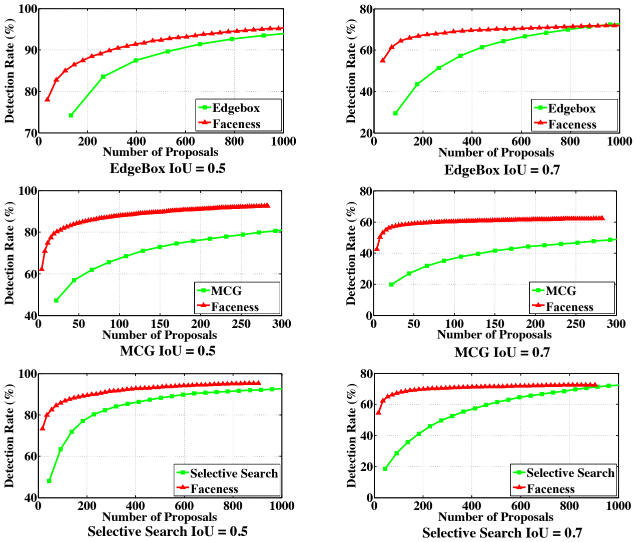
Table 3. The number of proposals needed for different recalls.

Proposal method	75%	80%	85%	90%
EdgeBox [41]	132	214	326	600
EdgeBox [41]+Faceness	**21**	**47**	**99**	**288**
MCG [2]	191	292	453	942
MCG [2]+Faceness	**13**	**23**	**55**	**158**
Selective Search [32]	153	228	366	641
Selective Search [32]+Faceness	**24**	**41**	**91**	**237**

Evaluate the contribution of each face part. We factor the contributions of different face parts to face proposal. Specifically, we generate face proposals with partness maps from each face part individually using the same evaluation protocol in previous experiment. As can be observed from Fig. 8(a), the hair, eye, and nose parts perform much better than mouth and beard. The lower part of the face is often occluded, making the mouth and beard less effective in proposing face windows. In contrast, hair, eye, and nose are visible in most cases. Nonetheless, mouth and beard can provide complementary cues.

Face proposals with different training strategies. As discussed in Sec. 3.1, there are different fine-tuning strategies that can be considered for generating a response map. We

[1] The face part bounding box is generated by first conducting non-maximum suppression (NMS) on the partness maps, and finding bounding boxes centered on NMS points.



EdgeBox IoU = 0.5

EdgeBox IoU = 0.7

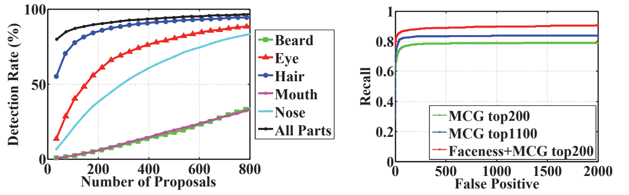
MCG IoU = 0.5

MCG IoU = 0.7

Selective Search IoU = 0.5

Selective Search IoU = 0.7

Figure 7. Comparing the performance between the proposed faceness measure and various generic objectness measures on proposing face candidate windows.



(a)

(b)

Figure 8. (a) Contribution of different face parts on face proposal. (b) FDDB face detection results with different proposal methods.

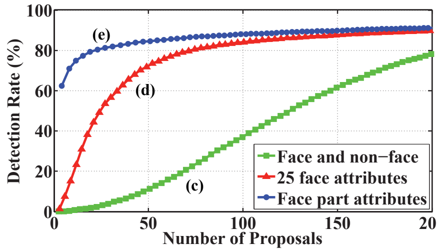


Figure 9. Comparing face proposal performance between different training strategies. Methods (c)-(e) are similar to those in Fig. 4. Method (e) is our approach.

compare face proposal performance between different training strategies. Quantitative results in Fig. 9 shows that our approach performs significantly better than approaches (c) and (d). This suggests that attributes-driven fine-tuning is more effective than 'face and non-face' supervision. As can be observed in Fig. 4 our method generates strong response even on the occluded face compared with approach (d), which leads to higher quality of face proposal.

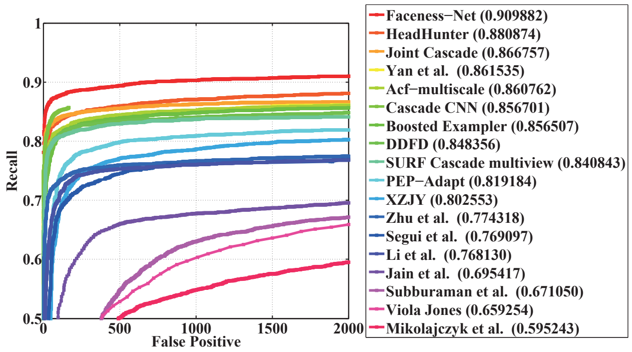


Figure 10. FDDB results. Recall rate is shown in the parenthesis.

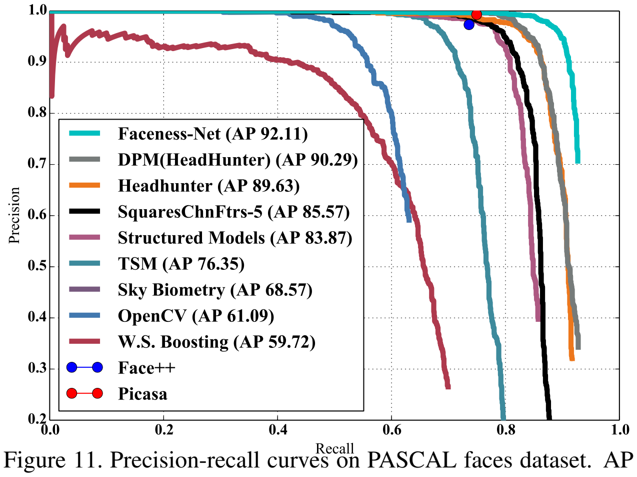


Figure 11. Precision-recall curves on PASCAL faces dataset. AP = average precision.

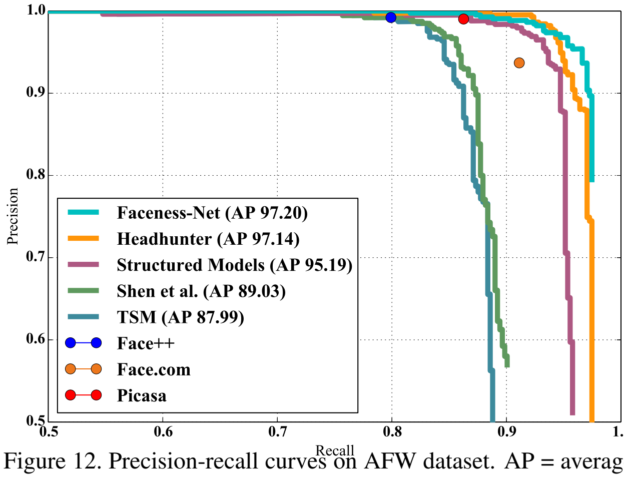


Figure 12. Precision-recall curves on AFW dataset. AP = average precision.

5.3. From Face Proposal to Face Detection

In this experiment, we first show the influence of training a face detector using generic object proposals and our face proposals. Next we compare our face detector, Faceness-Net, with state-of-the-art face detection approaches.

Generic object proposal versus face proposal. We choose the best performer in Fig. 7, *i.e.* MCG, to conduct this com-

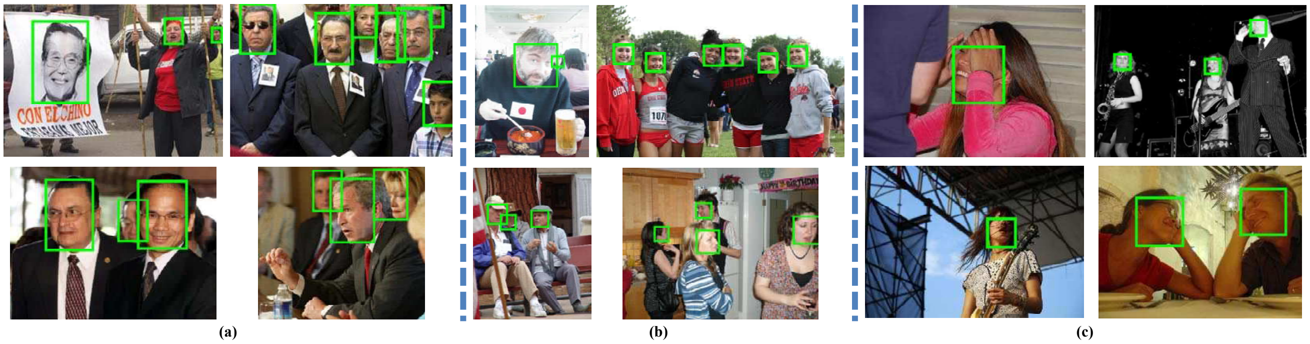


Figure 13. Qualitative face detection results by Faceness-Net on FDDB (a), AFW (b), PASCAL faces (c).

parison. The result is shown in Fig. 8(b). The best performance, a recall of 93%, is achieved by using our faceness measure to re-rank the MCG top 200 proposals (Faceness+MCG top-200). Using MCG top 200 proposals alone yields the worst result. Even if we adjust the number of MCG proposal to $1,100$ with a high recall rate similar to that of our method, the result is still inferior due to the enormous number of false positives. The results suggest that the face proposal generated by our approach is more accurate in finding faces than generic object proposals for face detection.

Comparison with face detectors. We conduct face detection experiment on three datasets FDDB [12], AFW [40] and PASCAL faces [36]. Our face detector, Faceness-Net, is trained with top 200 proposals by re-ranking MCG proposals following the process described in Sec. 3.3. We adopt the PASCAL VOC precision-recall protocol for evaluation.

We compare Faceness-Net against all published methods [37, 23, 3, 35, 18, 20, 17, 28, 40, 13] in the FDDB. For the PASCAL faces and AFW we compare with (1) deformable part based methods, *e.g.* structure model [36] and Tree Parts Model (TSM) [40]; (2) cascade-based methods, *e.g.* Headhunter [23]. Figures 10, 11, and 12 show that Faceness-Net outperforms all previous approaches by a considerable margin, especially on the FDDB dataset. Fig 6(b) shows some qualitative results on FDDB dataset together with the partness maps. More detection results are shown in Fig 13.

6. Discussion

There is a recent and concurrent study that proposed a Cascade-CNN [19] for face detection. Our method differs significantly to this method in that we explicitly handle partial occlusion by inferring face likeliness through part responses. This difference leads to a significant margin of 2.65% in recall rate (Cascade-CNN 85.67%, our method 88.32%) when the number of false positives is fixed at 167 on the FDDB dataset. The complete recall rate of the proposed Faceness-Net is 90.99% compared to 85.67% of Cascade-CNN.

At the expense of recall rate, the fast version of Cascade-CNN achieves 14fps on CPU and 100fps on GPU for 640×480 VGA images. The fast version of the proposed Faceness-Net can also achieve practical runtime efficiency, but still with a higher recall rate than the Cascade-CNN. The speed up of our method is achieved in two ways. First, we share the layers from conv1 to conv5 in the first stage of our model since the face part responses are only captured in layer conv7 (Fig. 2). The computations below conv7 in the ensemble are mostly redundant, since their filters capture global information *e.g.* edges and regions. Second, to achieve further efficiency, we replace MCG with Edgebox for faster generic object proposal, and reduce the number of proposal to 150 per image. Under this aggressive setting, our method still achieves a 87% recall rate on FDDB, higher than the 85.67% achieved by the full Cascade-CNN. The new runtime of our two-stage model is 50ms on a single GPU[2] for VGA images. The runtime speed of our method is comparatively lower than [19] because our implementation is currently based on unoptimized MATLAB code.

We note that further speed-up is possible without much trade-off on detection performance. Specifically, our method will benefit from Jaderberg *et al.* [11], who show that a CNN structure can enjoy a 2.5× speedup with no loss in accuracy by approximating non-linear filtering with low-rank expansions. Our method will also benefit from the recent model compression technique [8].

Acknowledgement This work[3] is partially supported by the National Natural Science Foundation of China (91320101, 61472410, 61503366).

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *TPAMI*, 2012. 3

[2] We use the same Nvidia Titan Black GPU as in Cascade-CNN [19].

[3] For more technical details, please contact the corresponding author Ping Luo via pluo.lhi@gmail.com.

[2] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *CVPR*, 2014. 3, 4, 6

[3] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun. Joint cascade face detection and alignment. In *ECCV*. 2014. 1, 2, 8

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 5

[5] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *TPAMI*, 2013. 4

[6] S. S. Farfade, M. Saberian, and L.-J. Li. Multi-view face detection using deep convolutional neural networks. *arXiv*, 2015. 2

[7] C. Garcia and M. Delakis. Convolutional face finder: A neural architecture for fast and robust face detection. *TPAMI*, 2004. 1, 2

[8] G. E. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *CoRR*, 2015. 8

[9] C. Huang, H. Ai, Y. Li, and S. Lao. High-performance rotation invariant multiview face detection. *TPAMI*, 2007. 1, 2

[10] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007. 5

[11] M. Jaderberg, A. Vedaldi, and A. Zisserman. Speeding up convolutional neural networks with low rank expansions. In *BMVC*, 2014. 8

[12] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical report, University of Massachusetts, Amherst, 2010. 5, 8

[13] V. Jain and E. Learned-Miller. Online domain adaptation of a pre-trained cascade of classifiers. In *CVPR*, 2011. 8

[14] A. Kae, K. Sohn, H. Lee, and E. Learned-Miller. Augmenting crfs with boltzmann machine shape priors for image labeling. In *CVPR*, 2013. 5, 6

[15] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011. 5

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 3, 4, 5

[17] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic part model for unsupervised face detector adaptation. In *ICCV*, 2013. 8

[18] H. Li, Z. Lin, J. Brandt, X. Shen, and G. Hua. Efficient boosted exemplar-based face detection. In *CVPR*, 2014. 8

[19] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *CVPR*, 2015. 8

[20] J. Li and Y. Zhang. Learning surf cascade for fast and accurate object detection. In *CVPR*, 2013. 1, 2, 8

[21] Y.-Y. Lin and T.-L. Liu. Robust face detection with multi-class boosting. In *CVPR*, 2005. 2

[22] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh. Fast object detection with occlusions. In *ECCV*, 2004. 2

[23] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *ECCV*. 2014. 1, 2, 8

[24] V. Mnih and G. Hinton. Learning to label aerial images from noisy data. In *ICML*, 2012. 4

[25] M. Osadchy, Y. Le Cun, and M. L. Miller. Synergistic face detection and pose estimation with energy-based models. *JMLR*, 2007. 1, 2

[26] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *TPAMI*, 1998. 1, 2

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. ImageNet large scale visual recognition challenge. *arXiv*, 2014. 4

[28] X. Shen, Z. Lin, J. Brandt, and Y. Wu. Detecting and aligning faces by image retrieval. In *CVPR*, 2013. 8

[29] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv*, 2013. 3

[30] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *NIPS*, 2014. 4

[31] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *CVPR*, 2014. 5

[32] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *IJCV*, 2013. 3, 4, 6

[33] R. Vaillant, C. Monrocq, and Y. Le Cun. Original approach for the localisation of objects in images. *VISP*, 1994. 1, 2

[34] P. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 2004. 1, 2

[35] J. Yan, Z. Lei, L. Wen, and S. Li. The fastest deformable part model for object detection. In *CVPR*, 2014. 8

[36] J. Yan, X. Zhang, Z. Lei, and S. Z. Li. Face detection by structural models. *IVC*, 2014. 1, 2, 5, 8

[37] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Aggregate channel features for multi-view face detection. *CoRR*, 2014. 8

[38] M.-H. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *TPAMI*, 2002. 2

[39] M. Zeiler and R. Fergus. Visualizing and understanding convolutional neural networks. In *ECCV*, 2014. 3

[40] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012. 1, 2, 5, 8

[41] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 3, 4, 5, 6