

Faceness-Net: Face Detection through Deep Facial Part Responses

Shuo Yang, Ping Luo, Chen Change Loy, *Member, IEEE* and Xiaou Tang, *Fellow, IEEE*

Abstract—We propose a deep convolutional neural network (CNN) for face detection leveraging on facial attributes based supervision. We observe a phenomenon that part detectors emerge within CNN trained to classify attributes from uncropped face images, without any explicit part supervision. The observation motivates a new method for finding faces through scoring facial parts responses by their spatial structure and arrangement. The scoring mechanism is data-driven, and 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 variations. Our method achieves promising performance on popular benchmarks including FDDB, PASCAL Faces, AFW, and WIDER FACE.

Index Terms—Face Detection, Deep Learning, Convolutional Neural Network.

arXiv:1701.08393v1 [cs.CV] 29 Jan 2017

1 INTRODUCTION

Face detection is an important and long-standing problem in computer vision. A number of methods have been proposed in the past, including neural network based methods [1], [2], [3], [4], cascade structures [5], [6], [7], [8] and deformable part models (DPM) [9], [10], [11] detectors. There has been a resurgence of interest in applying convolutional neural networks (CNN) on this classic problem [12], [13], [14], [15]. Many of these methods follow a cascade object detection framework [16], some of which directly adopt the effective generic object detection framework Faster-R-CNN [17] as the backbone network, with very deep networks (*e.g.*, 101-layer ResNet) to leverage the remarkable representation learning capacity of deep CNN [15].

While face bounding boxes have been used as a standard supervisory source for learning a face detector, the usefulness of facial attributes remains little explored. In this study, we show that facial attributes based supervision can effectively enhance the capability of a face detection network in handling severe occlusions. As depicted in Fig. 1, a CNN supervised with facial attributes can detect faces even when more than half of the face region is occluded. In addition, the CNN is capable of detecting faces with large pose variation, *e.g.*, profile view without training separate models under different viewpoints. Such compelling results are hard to achieve by using supervision based on face bounding boxes alone, especially when the training dataset has limited scene diversity and pose variations.

In this study, we show the benefits of facial attributes supervision through the following considerations:

(1) *Discovering facial parts responses supervised by facial attributes*: Human face has a unique structure. We believe the reasoning of the unique structure of local facial parts (*e.g.*, eyes, nose, mouth) help detecting faces under unconstrained environments. We observe an interesting phenomenon that one can actually obtain part detectors within a CNN by training it to classify part-level binary attributes (*e.g.*, mouth attributes including big lips, opened mouth, smiling, wearing lipstick) from uncropped face

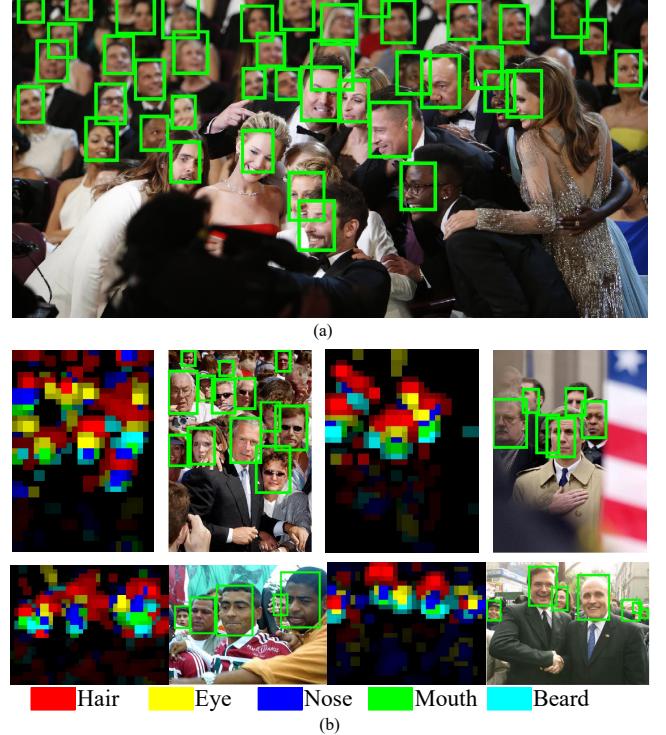


Fig. 1. (a) We propose a deep convolutional network for face detection, which achieves high recall of faces even under severe occlusions and head pose variations. The key to the success of our approach is the new mechanism for scoring face likeliness based on deep network responses on local facial parts. (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 approaches.

images, without any explicit part supervision. The trained CNN is then capable of generating high-quality facial part responses in its deep layers that strongly indicate the locations of the face parts. The examples depicted in Fig. 1(b) show the response maps (known as ‘partness map’ in our paper) of five different face parts.

• Department of Information Engineering, The Chinese University of Hong Kong.
E-mail: {ys014, pluo, ccloy, xtang}@ie.cuhk.edu.hk

(2) *Computing faceness score from responses configurations:* Given the parts’ responses, we formulate an effective method to reason the degree of face likeliness (which we call *faceness score*) through analyzing 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 windows¹ to obtain a set of face proposals. Our face proposal approach enjoys a high recall with just a modest number of proposals (over 90% of face recall with around 150 proposals, $\approx 0.5\%$ of full sliding windows, and $\approx 10\%$ of generic object proposals [18], measured on the FDDB dataset [19]).

(3) *Refining the face hypotheses* – Both the aforementioned components offer a 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 multi-task CNN 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 CNN and attributes supervision for discovering facial parts’ responses. We show that part detectors emerge within a CNN trained to classify attributes from uncropped face images, without any explicit part supervision. The parts’ responses are subsequently employed to generate high-quality proposals for training a face detector that is robust to severe occlusion. The findings aforementioned are new in the literature. It is worth pointing out that our network is trained on datasets that are not targeted for face detection (CelebA [20] for face recognition, and AFLW [21] for face alignment) and with simple background. Nevertheless, it still achieves promising performance on various face detection benchmarks including FDDB, PASCAL Faces, AFW, and the challenging WIDER FACE dataset.

In comparison to our earlier version of this work [22], we present a more effective design of CNN to achieve improved performance and speed. Firstly, in contrast to our previous work that requires independent convolutional networks for learning responses of different facial parts, we now share feature representations between these attribute-aware networks. The sharing of low and mid-levels representations largely reduce the number of parameters in our framework ($\sim 83\%$ fewer parameters), while improving the robustness of the feature representation. Secondly, our previous framework relies on external generic object proposal generators such as selective search [23] and EdgeBox [24] for proposing candidate windows. Inspired by region proposal network presented in [17], in this study we directly generate proposals from our attribute-aware networks, thus proposal generation becomes an inherent part of the framework. This design not only leads to improved computation efficiency but also higher recall rate compared with generic object proposal algorithms. Thirdly, we compare our face detector pre-trained on the task of facial attributes classification with that pre-trained on ImageNet large-scale object classification. Apart from the above major changes, we also provide more technical details and discussions. Additional experiments are conducted on the challenging WIDER FACE dataset [25].

1. There are many options to generate candidate windows. We show two options in this study: (i) using generic object proposal generator, and (ii) using a region proposal network. See Sec. 3.3 for details.

2 RELATED WORK

There is a long history of using neural network for the task of face detection [1], [2], [3], [4]. An early face detection survey [26] provides an extensive coverage on relevant methods. Here we highlight a few notable studies. Rowley *et al.* [2] 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.* [4] 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.* [1] 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. For instance, our first stage network is conceptually different from that of [1], and many recent deep learning detection frameworks – we train attribute-aware 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 [27], [28], 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 [5], [6], [7], [8] and deformable part models (DPM) detectors dominate face detection approaches. Viola and Jones [8] introduced fast Haar-like features computation via integral image and boosted cascade classifier. Various studies thereafter follow a similar pipeline. Among the variants, SURF cascade [7] was one of the top performers. Later Chen *et al.* [5] 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 [9] demonstrates state-of-the-art performance with just a vanilla DPM, achieving better results than more sophisticated DPM variants [10], [11].

Recent studies [13], [16], [29], [30], [31] show that face detection can be further improved by using deep learning. The network proposed by [29] does not have an explicit mechanism to handle occlusion, the face detector therefore fails to detect faces with heavy occlusions, as acknowledged by the authors. Cascade based convolutional neural networks [12], [16] replace boosting classifiers with a set of small CNNs to quickly reject negative samples in the early stage. Recent studies [13], [30] exploit facial landmarks as supervision signals to improve face detection performance. In this study, we show that facial attributes can serve as an important source too for learning a robust face detector.

The first stage of our model is partially inspired by generic object proposal approaches [23], [32], [33]. Generic object proposal generators are commonly used in standard object detection algorithms for providing high-quality and category-independent bounding boxes. These methods typically involve redundant computations over regions that are covered by multiple proposals. To reduce computation, Ren *et al.* [17] propose Regional Proposal Network (RPN) to generate proposals from high-level response maps in a CNN through a set of predefined anchor boxes. Both

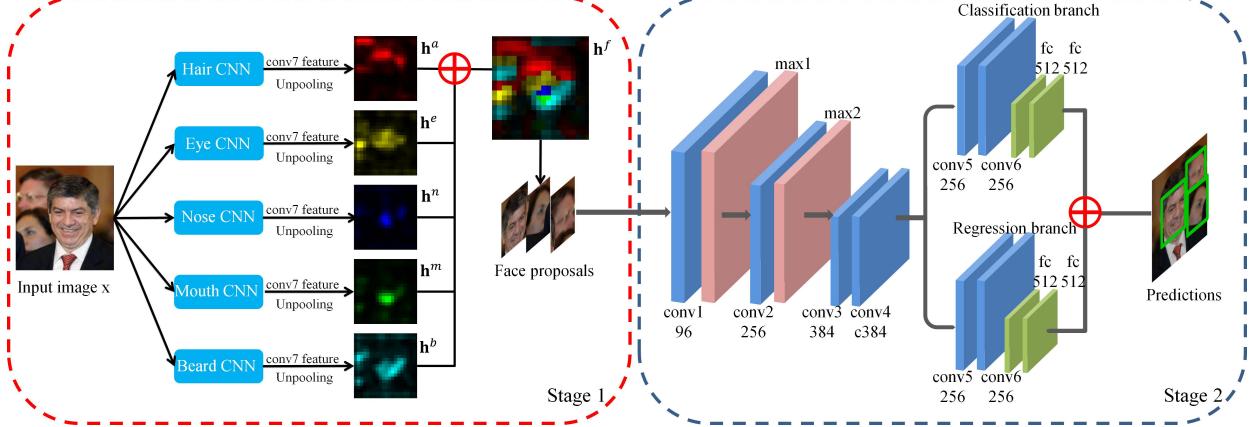


Fig. 2. The pipeline of the baseline *Faceness-Net*. The first stage of *Faceness-Net* applies attribute-aware networks to generate response maps of different facial parts. The maps are subsequently employed to produce face proposals. The second stage of *Faceness-Net* refines candidate window generated from first stage using a multi-task convolutional neural network (CNN), where face classification and bounding box regression are jointly optimized. (Best viewed in color).

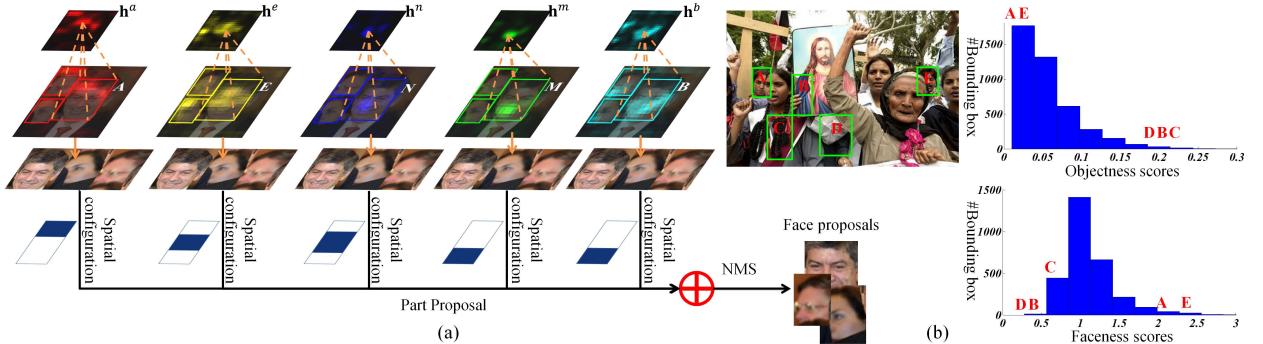


Fig. 3. (a) The pipeline for generating face proposals. (b) Bounding box re-ranking by face measure (Best viewed in color).

generic object proposal and RPN methods do not consider the unique structure and parts on the face. Hence, no principled mechanism is available 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, while reducing the number of candidate windows to half the original (compared to the original RPN [17]).

3 FACENESS-NET

This section introduces the baseline Faceness-Net. We first briefly overview the entire pipeline and then discuss the details. As shown in Fig. 2, Faceness-Net consists of two stages, *i.e.*, (i) generating face proposals from partness maps by ranking candidate windows using faceness scores, and (ii) refining face proposals for face detection.

First stage. A full image \mathbf{x} is used as an input to a CNN to generate the partness map for each face part. A set of CNNs, known as attribute-aware networks, are used to generate the partness map of different parts independently. The partness map is obtained by weighted averaging over all the response maps at its top convolutional layer. The map 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. For illustration, we add all these maps into a face label map \mathbf{h}^f , which clearly suggests faces' locations

Given a set of candidate windows $\{w\}$ that are generated by existing object proposal methods such as [23], [32], [33], or a region proposal network (RPN) [17], we rank these windows according to their faceness scores, Δ_w , which are derived from the partness maps with respect to different facial parts configurations, as illustrated at the bottom of Fig. 3(a). For example, as visualized in Fig. 3(a), 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. The bottom part of Fig. 3(a) illustrates the spatial configurations of five facial parts. The facial configurations can be learned from the training data. 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. To further reduce the number of the proposed windows, we apply non-maximum suppression (NMS) to smooth the scores by leveraging the spatial relations among these windows. The proposed approach is capable of coping with severe face occlusions, as shown in Fig. 3(b), where face windows 'A' and 'E' can be retrieved by objectness [34] only if large amount of windows are proposed, while they rank top 50 by using our method.

Second stage. The face proposals are refined by training a multi-task CNN, where face classification and bounding box regression are jointly optimized (Fig. 2).

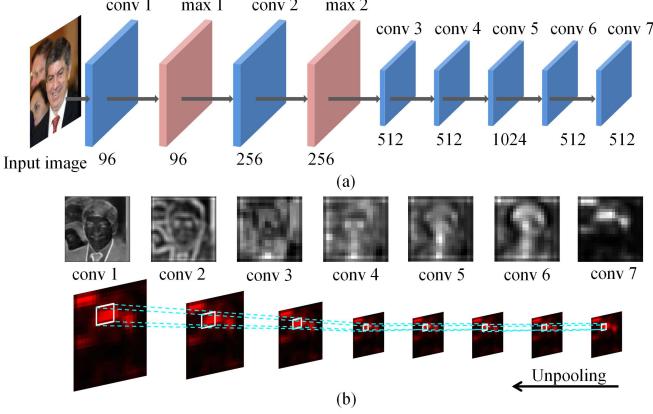


Fig. 4. In the baseline Faceness-Net, we adopt different attribute-aware networks for different facial parts. (a) This figure shows the architecture of an attribute-aware deep network used for discovering the responses of ‘hair’ component. Other architectures are possible. See Sec. 3.1 for details. (b) The average-pooled response maps from Conv7 indicate the location of hair component. The response map is upsampled through unpooling operation [36] to obtain the final map of the same size as the input image.

3.1 Attribute-Aware Networks

The first stage of the baseline Faceness-Net consists of multiple attribute-aware networks for generating response maps of different parts (Fig. 2). Five networks are needed to cover all five pre-defined facial components, *i.e.*, hair, eyes, nose, mouth, and beard. These attribute-aware networks share the same structure. Next, we first discuss the network structure and subsequently show that these networks can share representation to save parameters.

Network structure. The choice of network structure for extracting partness maps is flexible. Figure 4(a) depicts the structure and hyper-parameters of the CNN used in the baseline Faceness-Net. This convolutional structure is inspired by the AlexNet [35], which is proposed for object categorization. Specifically, the network stacks seven convolutional layers (Conv1 to Conv7) and two max-pooling layers (Max1 and Max2). The hyper-parameters of each layer is specified in Fig. 4(a).

Once the attribute networks are trained (training details are provided in Sec. 3.2), we examine the response maps obtained from the attribute-aware networks. As observed from Fig. 4(b), the feature maps of the first few convolutional layers do not clearly indicate the locations of facial parts. However, clear indication of the facial component can be seen from the average-pooled responses of Conv7. Consequently, we obtain the initial partness map from the output layer, which contains a single filter with a size of $1 \times 1 \times 512$. The role of this filter is to perform weighted average on the feature maps of Conv7. The final partness map that matches the input image’s size is obtained through unpooling operation [36].

Shared representation. It is observed that the feature maps of earlier layers across the different attribute-aware networks are almost identical and they are not indicative of parts’ locations. Motivated by these observations, instead of designating separate attribute-aware networks for different facial components, we share early convolutional layers of these networks to reduce parameters. Specifically, the first four convolutional layers that do not clearly suggest parts’ locations are shared, followed by five branches, each of which consists of two convolutional layers responsible for

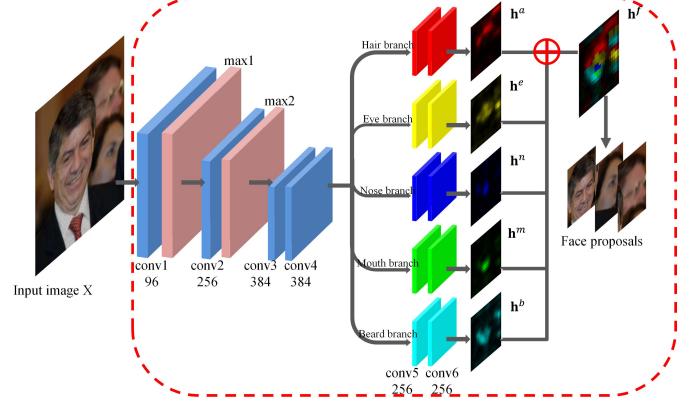


Fig. 5. The first stage of Faceness-Net-SR, a variant of the baseline Faceness-Net. We share representations between the different attribute-aware networks and reduce filters leading to improved efficiency and performance. See Sec. 3.1 for details.

a facial component, as shown in Fig. 5. Note that in comparison to the structure presented in Fig. 4(a), we additionally remove a convolutional layer and trim the number of filters in other layers to reduce parameters. The sharing of representation and filter reduction lead to a single attribute-aware network with 83% fewer parameters than the original five attribute-aware networks. We denote a Faceness-Net with shared representation as Faceness-Net-SR. We will show that this network structure not only reduces computations, but also improves the robustness of feature representation for face detection.

3.2 Learning to Generate Partness Maps

Pre-training the attribute-aware networks. Pre-training generally helps improving the performance of a deep network. There are two plausible pre-training options depending upon whether we share the representations across attribute-aware networks or not.

The first option is to pre-train our attribute-aware networks with massive general object categories in ImageNet [37]. From our observations, this option works well when the representations across networks are not shared. Since each attribute-aware network originally has access only to a particular group of data specific to a certain attribute, the larger-scale ImageNet data helps to mitigate the overfitting issue that is caused by insufficient data.

The second option omits the ImageNet pre-training stage and directly trains a network on the task of facial attributes classification. This option works best when we adopt the shared representation scheme discussed in Sec. 3.1. Thanks to the sharing of representation, the attribute-aware network requires relatively smaller quantity of training data. Thus, no overfitting is observed despite we use the facial attributes dataset, which is much smaller in scale, *i.e.*, 180,000 images compared to 1 million images in ImageNet.

Fine-tuning the attribute-aware networks. Once an attribute-network is pre-trained, we can fine-tune it to generate the desired partness maps. There are different fine-tuning strategies, but not all of them can generate meaningful partness maps for deriving a robust faceness score.

As shown in Fig. 6(b), a deep network trained on generic objects, *e.g.*, AlexNet [35], is not capable of providing us with precise faces’ locations, let alone partness map. To generate accurate partness maps, we explore multiple ways for learning

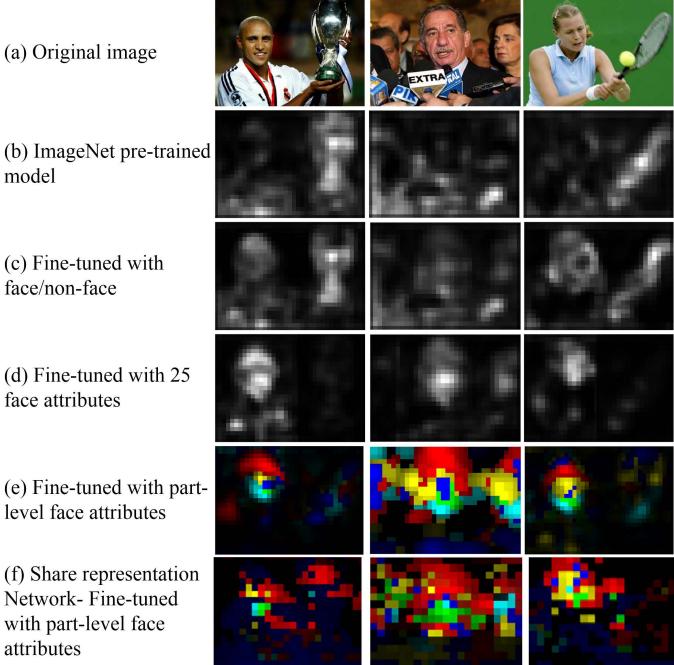


Fig. 6. The partness maps obtained by using different types of supervisions and fine-tuning strategies. The maps in (a-e) are generated using the baseline Faceness-Net depicted in Fig. 2. The maps in (f) is generated using Faceness-Net-SR with shared representation, as illustrated in Fig. 5.

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 |

an attribute-aware network. The most straight-forward manner is to use the image and its pixel-wise segmentation label map as input and target, respectively. This setting is widely employed in image labeling [38], [39]. 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. 6(c). It works well where the training images are well-aligned, such as face recognition [40]. Nevertheless, it suffers from complex background clutter because the supervisory information is not sufficient to account for rich and diverse face variations. Its learned feature maps contain too much noises, which overwhelm the actual faces' locations. Attribute learning in Fig. 6(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.

Figure 6(e) shows the partness maps obtained by the baseline FacenessNet, of which the attribute networks do not share representations. The strategy we propose 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, face parts are modeled separately. If one part is occluded, the face region can still be localized by the other parts. We take the Hair-Branch shown in the stage one of Fig. 2 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 $\forall \mathbf{x}_i \in \mathbb{R}^{128 \times 128}$ and $\forall \mathbf{y}_i \in \mathbb{R}^{1 \times 92}$, implying that each full image is rescaled to 128×128 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,

$$L = \sum_{i=1}^N \mathbf{y}_i \log p(\mathbf{y}_i = 1 | \mathbf{x}_i) + (1 - \mathbf{y}_i) \log (1 - p(\mathbf{y}_i = 1 | \mathbf{x}_i)), \quad (1)$$

where $p(\mathbf{y}_i | \mathbf{x}_i)$ is modeled as a sigmoid function, *i.e.* $p(\mathbf{y}_i = j | \mathbf{x}_i) = \frac{1}{1 + \exp(-f(\mathbf{x}_i))}$, indicating the probability of the presence of the attributes. The features of \mathbf{x}_i is denoted as $f(\mathbf{x}_i)$ and $\mathbf{1}$ represents a vector with all elements equal one. To facilitate the learning, we stack two fully-connected layers on top of the last convolutional layer of the structure shown in Fig. 4. After which we optimize the loss function by using the stochastic gradient descent with back-propagation. After training the attribute-aware network, the fully-connected layers are removed to make the network fully convolutional again.

Figure 6(f) shows the partness maps that are generated from the networks with shared representation, *i.e.*, Faceness-Net-SR (see Fig. 5). Visually, the partness maps generated by this model are more noisy compared to Fig. 6(e). The key reason is that the Faceness-Net-SR is not pre-trained using ImageNet data but directly trained on the attribute classification task. Despite the noisy partness maps, they actually capture more subtle parts' responses and therefore lead to higher recall rate in the subsequent face proposal stage, provided that the number of proposals is sufficiently large.

3.3 Generating Candidate Windows

Face detection can be improved if the inputs are formed by moderate number of proposals with a high-recall rate. To produce the required proposals, we will explore two plausible choices to generate the initial set of candidate windows.

Generic object proposal. Generic object scoring is primarily employed to reduce the computational cost of a detector. It has also been shown improving detection accuracy due to reduction of spurious false positives [34]. A variety of cues have been proposed to quantify the objectness of an image window, *e.g.*, norm of the gradient [42], edges [33], or integration of a number of low-level features [34]. Other popular methods include super-pixel based approaches, *e.g.*, selective search [23], randomized Prim [43], and multi-scale combinatorial grouping [32]. Our framework can readily employ these generic candidate windows for ranking using the proposed faceness score (Sec. 3.4).

Region proposal network. Previous study [17] employs deep neural network to generate region proposals through conducting object/non-object classification and bounding box regression. We adapt this idea to our framework with a few changes.

2. Other target designs [41] are applicable.

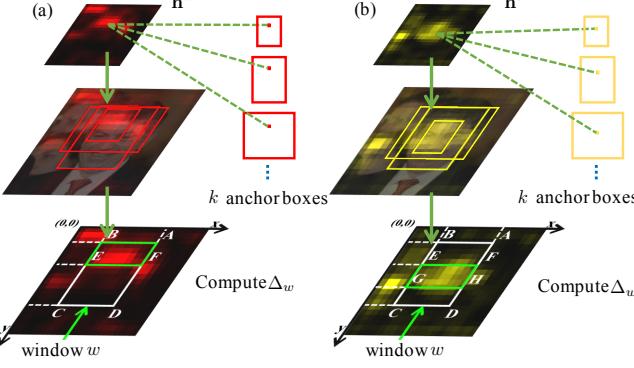


Fig. 7. Examples of region proposal and faceness measurement. The partness maps of hair and eyes are shown in (a) and (b), respectively. Δ_w is the faceness score of a window, w . (Best viewed in color).

- (a) *Proposing regions on partness maps* - instead of using the generic object feature maps, we exploit the partness maps produced by face attribute-aware networks for proposing the initial candidate windows.
- (b) *Omitting face/non-face classification and bounding box regression* - Different from RPN [17], we skip the learning from face/non-face classification and bounding box regression at this stage. The partness maps that are learned from face attributes classification already provide high-quality response maps to propose face candidate windows.

We provide an example below on using a partness map of hair, \mathbf{h}^a , for region proposal. As shown in Fig. 7, each value of location (i, j) on the partness map \mathbf{h}^a indicates the probability of the appearance of the hair component. We select a set of M locations $\{(h_i, h_j)\}_{i=1}^M$ with a probability $p(h_i, h_j)$ higher than t . For each selected location, multiple region proposals are generated, where the number of maximum possible proposals for each location is fixed as k . The proposals are obtained from predefined reference boxes, which we call anchors. For each face part, anchors are centered at different locations considering the structure of human face. In addition, they are associated with a specific scale and aspect ratio, as shown at the top of Fig. 7. For instance, the anchors of the hair region are centered at $(W/2, H/3)$ and the anchors of eyes are centered at $(W/2, H/2)$, where W and H represent the width and height of an anchor. Similar to previous work [17], these anchors are translation invariant up to the network's total stride, and the method does not incur extra cost for addressing scales thanks to the multi-scale anchors.

In our study, we define 10 scales and 1 aspect ratio, yielding $k = 10$ anchors at each selected position. Specifically, we use 10 scales with box areas of $25^2, 50^2, 75^2, 100^2, 135^2, 170^2, 200^2, 240^2, 300^2$, and 350^2 pixels, and 1 aspect ratio of $1 : 1.5$ (with width to height). The parameters of anchors, *i.e.*, center location, scale and aspect ratio are selected by maximizing the recall rate given an average number of n proposal per image. In our study, we perform grid search on the training set to select the parameters.

Discussion. Both the generic objectness measures and original RPN (trained on ImageNet) 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. While RPN is computationally more efficient than generic object proposal generators, it cannot be directly applied to

our problem too. Specifically, in order for the RPN to cope with faces with tiny size and various poses, a large number of anchor boxes are required, leading to enormous number of proposals. In the next section, we discuss a new faceness measure that can complement existing object proposal generators to achieve high recall on faces, while significantly reduce the number of candidate windows.

3.4 Ranking Windows by Faceness Score

After generating candidate windows based on the methods described in Sec. 3.3, our approach computes faceness score on these windows to return a ranked set of top-scoring face proposals. Figure 7 illustrates the procedure of deriving the faceness measure from the partness maps of hair and eyes. Let Δ_w be the faceness score of a window w . For example, as shown in Fig. 7(a), given a partness map of hair, \mathbf{h}^a , Δ_w is attained by dividing the sum of values in ABEF (green) by the sum of values in FECD. Similarly, Fig. 7(b) shows that Δ_w is obtained by dividing the sum of values in EFGH (green) with respect to ABEG+HGCD of \mathbf{h}^e . For both of the above examples, a larger value of Δ_w indicates a higher overlapping ratio of w with a face.

The spatial configurations, such as ABEF in Fig. 7(a) and EFGH in Fig. 7(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 Fig. 7(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

$$\frac{\mathbf{I}(x_f, y_f) + \mathbf{I}(x_b, y_b) - \mathbf{I}(x_a, y_a) - \mathbf{I}(x_b, \lambda y_b + (1 - \lambda) y_c)}{\mathbf{I}(x_d, y_d) + \mathbf{I}(x_e, y_e) - \mathbf{I}(x_a, \lambda y_a + (1 - \lambda) y_d) - \mathbf{I}(x_c, y_c)}, \quad (2)$$

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. Let \mathbf{h}_i be the cropped partness map with respect to the i -th window, *e.g.*, region ABCD in \mathbf{h}^a . This problem can be formulated as maximum a posteriori (MAP) estimation

$$\lambda^* = \operatorname{argmax}_{\lambda} \prod_i^M p(r_i | \lambda, w_i, \mathbf{h}_i) p(\lambda, w_i, \mathbf{h}_i), \quad (3)$$

where λ represents a set of parameters when learning the spatial configuration of hair (Fig. 7(a)). The terms $p(r_i | \lambda, w_i, \mathbf{h}_i)$ and $p(\lambda, w_i, \mathbf{h}_i)$ denote 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 the 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 obtained in Sec. 3.2. Since λ typically has a low dimension (*e.g.*, one dimension of hair), it can be simply obtained by line search. Note that Eq. (3) can be easily extended to model more complex spatial configurations.

3.5 Face Detection

The top candidate windows that are ranked by faceness measure attain a high recall rate. These face proposals can be subsequently fed to the multi-task CNN at stage 2 of the proposed pipeline (Fig. 2) for face detection.

Pre-training. We directly use the earlier layers of attribute-aware networks (the stage-1 network with shared representation as shown in Fig. 5) up to Conv4 as the pre-trained model for the multi-task CNN of stage 2. After Conv4, as shown in Fig. 2, the multi-task CNN forks into two branches, each of which consists of two convolutional layers and two fully connected layers. The two branches are optimized to handle different tasks, namely face classification and bounding box regression, respectively.

It is worth pointing out that the multi-task CNN can be pre-trained on the ImageNet data, instead of reusing the parameters of the attribute-aware networks. Nevertheless, we found that the multi-task CNN converges much faster given the face attributes based pretrained model. Specifically, the attribute pretrained network only requires 45,000 iterations to converge during the face detection fine-tuning stage, in comparison to more than 200,000 iterations for the ImageNet pretrained network using the same mini-batch size. We conjecture that much less efforts are needed to transform the feature representations learned from the facial attribute classification task to the face detection task.

Multi-task fine-tuning. We fine-tune the first branch of the multi-task CNN for face classification and the second branch for bounding box regression. Fine-tuning is performed using the face proposals obtained from the previous step (Sec 3.4). For face classification, we assign a face proposal to its closest ground truth bounding box based on the Euclidean distance between their respective center coordinates. A face proposal is considered positive if the Intersection over Union (IoU) between the proposal box and the assigned ground truth box is larger than 0.5; otherwise it is negative. For bounding box regression, we train the second branch of the multi-task CNN to regress each proposal to the coordinates of its assigned ground truth box. If the proposed window is a positive sample, the regression target is generated by Eq. (4). We use the following parameterizations of the 4 coordinates:

$$\begin{aligned} x_1^* &= (x_1 - x'_1)/\zeta, & y_1^* &= (y_1 - y'_1)/\zeta \\ x_2^* &= (x_2 - x'_1)/\zeta, & y_2^* &= (y_2 - y'_1)/\zeta, \end{aligned} \quad (4)$$

where $\zeta = \max(x'_2 - x'_1, y'_2 - y'_1)$ is a normalizing factor. The $[x_1, y_1, x_2, y_2]$ denotes the top-left and bottom-right coordinates of a bounding box. Variables x , x' , and x^* represent the ground truth box, proposed box, and regression target. If a proposed window is non-face, the CNN outputs a vector of $[-1, -1, -1, -1]$.

More implementation details are given below. During the training process, if the number of positive samples in a mini-batch is smaller than 20% of the total samples, we randomly crop the ground truth faces and add these samples as additional positive samples. Therefore, the ratio of positive samples and negative samples is kept not lower than 1 : 4. Meanwhile, we conduct bounding box NMS on the negative samples. The IoU for the NMS is set to 0.7. The proposed bounding boxes are cropped and then resized to 128×128 . To handle blurry faces, we augment our training samples by applying Gaussian blur. The fine-tuning consumes 50K iterations with a batch size of 256 images. We adopt Euclidean loss and cross-entropy loss for bounding box regression and face classification, respectively.

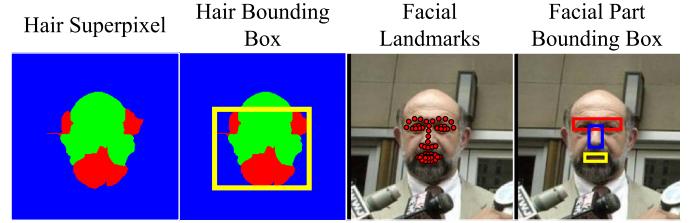


Fig. 8. Examples of ground truth bounding boxes of facial parts. Hair ground truth bounding boxes are generated from superpixel maps [46]. Eye, nose, and mouth bounding boxes are generated from 68 ground truth facial landmarks [47].

4 EXPERIMENTAL SETTINGS

Training datasets. (i) We employ CelebA dataset [20] to train our attribute-aware networks. The dataset contains 202,599 web-based images exclusive from the LFW [44], FDDB [19], AFLW [11] and PASCAL [10] datasets. Every image in the dataset are labeled with 40 facial attributes. We select 25 facial attributes from CelebA dataset for each image and divide the attributes into five categories based on their respective facial parts as shown in Table 1. We randomly select 180,000 images from the CelebA dataset for training and the remaining is reserved as the validation set. (ii) For face detection training, we choose 13,205 face images from the AFLW dataset [21] to ensure a balanced out-of-plane pose distribution. We observe a large number of missed annotated faces in the AFLW dataset, which could hamper the training of our face detector. Hence, we re-annotate face bounding boxes for those missing faces. The total number of faces in the re-annotated AFLW is 29,133 compared with 24,386 in the original data. As negative samples, we randomly select 5,771 person-free images from the PASCAL VOC 2007 dataset [45].

Part response test dataset. In Sec. 5.1, we use LFW dataset [44] for evaluating the quality of part response maps for part localization. Since the original dataset does not come with part-level bounding boxes, we label the boxes with the following scheme. We follow the annotations provided by [46] on hairs and beard for a set of 2927 LFW images. Hair bounding boxes are generated with minimal and maximal coordinates of hair superpixel as shown in Fig 8. Using similar strategy, eye, nose and mouth bounding boxes are obtained from the manually labeled 68 dense facial landmarks [47] on the original LFW [44] images, as shown in Fig 8.

Face proposal and detection test datasets. In Sec. 5.2 and Sec. 5.3, we use the following datasets. (i) FDDB [19] dataset contains 5,171 faces in a set of 2,845 images. For the face proposal evaluation, we follow the standard evaluation protocol widely used in object proposal studies [33] 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) AFLW [11] dataset contains 205 Flickr images with 473 annotated faces of large variations in both face viewpoint and appearance. (iii) PASCAL faces [10] is a widely used face detection benchmark dataset. It consists of 851 images and 1,341 annotated faces. (iv) WIDER FACE [25] is the largest and extremely challenging face detection benchmark dataset. It consists of 32,203 images and 393,703 annotated faces.

Evaluation settings. Following [33], we employ the Intersection over Union (IoU) as our evaluation metric. We fix the IoU thresh-

TABLE 2

Evaluating the robustness to unconstrained training input. Facial part detection rate is used. The number of proposals is 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% |

old to 0.5 following the strict PASCAL criterion. In particular, an object is considered being covered/detected by a proposal if the 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 [33]. For face detection, we use standard precision and recall (PR) to evaluate the effectiveness of face detection algorithms.

Baseline Faceness-Nets. We evaluate four variants of Faceness-Net:

- Faceness-Net - our baseline method with five attribute-aware networks Fig. 2. An external generic object proposal generator is adopted.
- Faceness-Net-SR - a variant with single attribute-aware network by sharing representations, as shown in Fig. 5. An external generic object proposal generator is adopted.
- Faceness-Net-RP - a variant of the Faceness-Net that adopts the region proposal technique to generate candidate windows.
- Faceness-Net-SR-RP - a variant of the Faceness-Net-SR that uses the region proposal technique to generate candidate windows.

The discussion on generic object proposal and region proposal techniques can be found in Sec. 3.3.

5 RESULTS

5.1 Evaluating the Attribute-Aware Networks

Robustness to unconstrained training input. The proposed attribute-aware networks do not assume well-cropped faces as input in both the training and test stages. To support this statement, we conduct an experiment by fine-tuning two attribute-aware networks as shown in the Fig. 4(a), each of which taking different inputs: (1) uncropped images, which may include large portion of background clutter apart the face; and (2) cropped images, which encompass roughly the face and shoulder regions. Some examples of uncropped images are shown in Fig. 18(a). The performance is measured based on the part detection rate³. Note that we combine the evaluation on ‘Hair+Beard’ to suit the ground truth provided by [46] (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

3. 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.

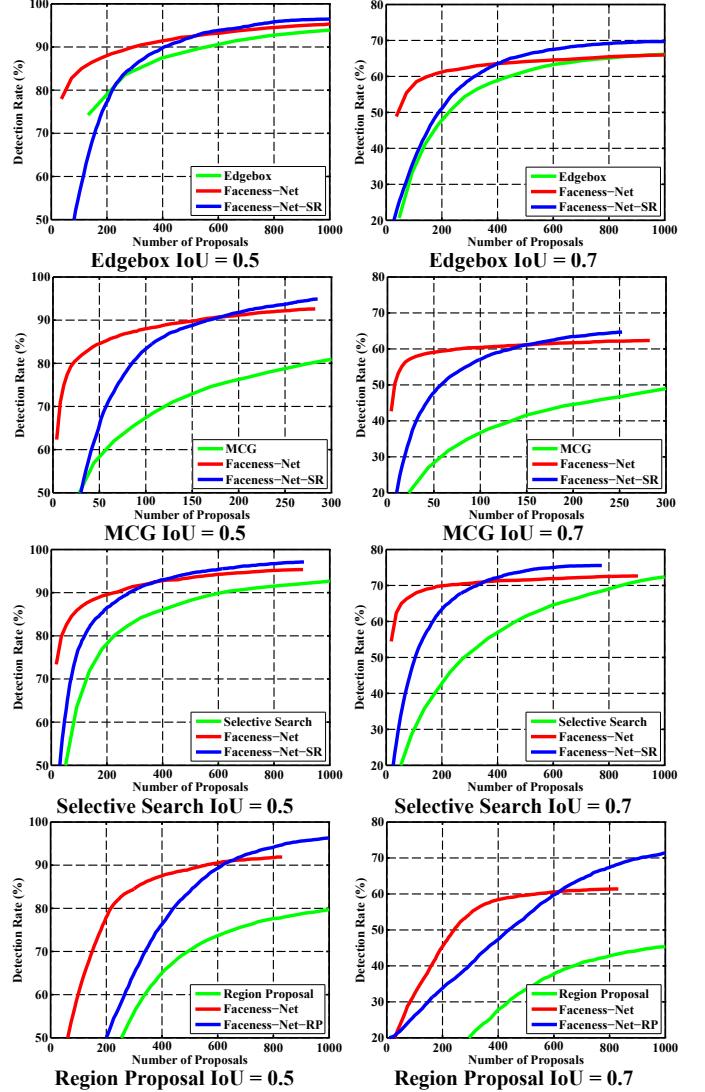


Fig. 9. The first three rows compare the performance between the proposed faceness measure and various generic objectness measures on proposing face candidate windows. The last row compares the performance between the proposed faceness measure and region proposal measures on proposing face candidate windows.

the following experiments, all the proposed models are trained on uncropped images. Fig. 18(a) shows the qualitative results. Note that facial parts can be discovered despite challenging poses.

With and without sharing representation. As mentioned in Sec. 3.1, we can train an attribute-aware network for each face part or we can train a single network for all the parts by sharing representation. We compare the proposal detection rate of these two options. Figure 9 shows the proposal detection rate of attribute-aware network(s) trained with and without sharing representation, indicated by blue and red curves, respectively. Attribute-aware networks trained without sharing representation require fewer number of proposals but with a detection rate typically lower than 90% (given 150-200 proposals). On the contrary, the attribute-aware network that shares low-level and mid-level representations can achieve a higher detection rate but with an expense of larger number of proposals.

The observations can be explained as follows. The networks

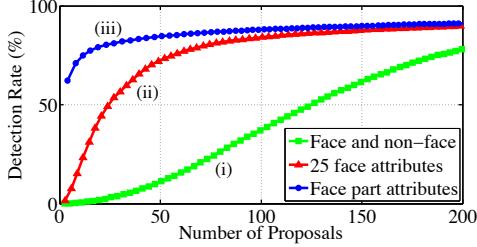


Fig. 10. Comparing the face proposal performance when different strategies are used to fine-tune the attribute-aware networks.

without sharing representation tend to model the discrepancies between individual face part and background, while the network that shares representation is more likely to learn the differences between facial parts. Thus, the latter has poorer background modelling capacity thus leading to inferior performance when the number of proposal is small, in comparison to the former. Nevertheless, we found that the network that shares representation yields high responses for subtle facial parts. This high recall rate is essential to improve the performance of face detection in the later stage.

Different fine-tuning strategies. As discussed in Sec. 3.2, there are different fine-tuning strategies that can be considered for learning to generate a partness map, but not all of them are well-suited for deriving a robust faceness measure. Qualitative results have been provided in Fig. 6. Here, we provide quantitative comparisons between the following fine-tuning approaches: (i) a network fine-tuned with a large number of face images from CelebA and non-face images, (ii) fine-tuning the network with 25 face attributes, and (iii) the proposed approach that fine-tunes attribute-aware networks with part-level attributes in accordance to Table 1. It is evident from Fig. 10 that our approach performs significantly better than approaches (i) and (ii).

5.2 From Part Responses to Face Proposal

Generic object proposal methods. In this experiment, we show the effectiveness of adapting different generic object proposal generators [23], [32], [33] 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 candidate windows and we use our faceness scoring method described in Sec. 3.4 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 [33].

The results are shown in the first three rows of Fig. 9. The green curves show the performance of baseline generic object proposal generators. It can be observed that our method consistently improves the baselines for proposing face candidate windows, under different IoU thresholds. Table 3 shows that our method achieves high detection rate with small number of proposals.

Region proposal method. The bottom row of Fig. 9 depicts the results of using region proposal (Sec. 3.3) to produce face-specific proposals. The proposed region proposal technique achieves a high detection rate of over 95% when the average number of proposal is 1,000 with an IoU of 0.5. It is worth pointing out that high quality region proposals are typically generated by a region proposal network that is trained by conducting object/non-object

TABLE 3
The number of proposals needed for different detection rate.

| Proposal method | 75% | 80% | 85% | 90% |
|-----------------------------------|------------|------------|------------|------------|
| EdgeBox [33] | 132 | 214 | 326 | 600 |
| EdgeBox [33]+Faceness | 21 | 47 | 99 | 288 |
| EdgeBox [33]+Faceness-SR | 180 | 210 | 275 | 380 |
| MCG [32] | 191 | 292 | 453 | 942 |
| MCG [32]+Faceness | 13 | 23 | 55 | 158 |
| MCG [32]+Faceness-SR | 69 | 87 | 112 | 158 |
| Selective Search [23] | 153 | 228 | 366 | 641 |
| Selective Search [23]+Faceness | 24 | 41 | 91 | 237 |
| Selective Search [23]+Faceness-SR | 94 | 125 | 189 | 309 |
| Region Proposal [17] | 649 | 997 | 1396 | 1732 |
| Region Proposal [17]+Faceness | 174 | 216 | 311 | 557 |
| Region Proposal [17]+Faceness-SR | 359 | 421 | 492 | 637 |

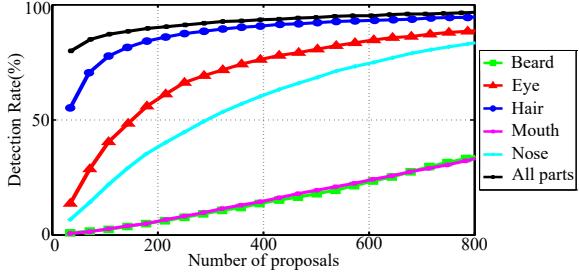


Fig. 11. The contributions of different face parts to face proposal.

classification and anchor box regression [14]. Our proposals, nonetheless, do not require such an explicit learning process (see Sec. 3.3), thanks to the high-quality responses of partness maps and faceness score.

Evaluate the contributions of each face part. We factor the contributions of different face parts to proposing face. 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. 11, 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. They therefore become important clues for face proposal. Nonetheless, mouth and beard could provide complementary cues. Thus combining all parts leads to better result than considering each part in isolation.

5.3 From Face Proposal to Face Detection

Next we compare the proposed Faceness-Net and its variants against state-of-the-art face detection approaches on four benchmark datasets including FDDB, AFW, PASCAL Faces, and WIDER FACE.

We conduct extensive face detection experiments on four benchmark datasets FDDB [19], AFW [11], PASCAL faces [10] and WIDER FACE [25]. Our baseline face detector, Faceness-Net, which involves five CNNs with the structure shown in the Fig. 2, is trained with top 200 proposals by re-ranking MCG proposals following the process described in Sec. 3.4. To factor the contributions of share representation and region proposal, we build another three variants of Faceness-Net as discussed in Sec. 4. The variant Faceness-Net-RP is trained with top 1000 region proposals that are re-ranked following Sec. 3.4.

We compare Faceness-Net and its variants against representative published methods [5], [7], [9], [11], [48], [49], [50], [51],

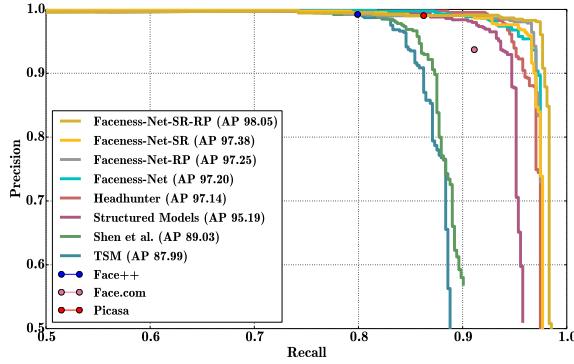


Fig. 12. Precision-recall curves on the AFW dataset. AP = average precision.

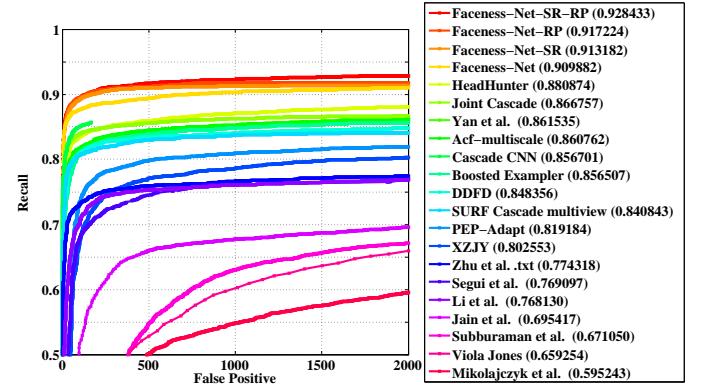


Fig. 14. FDDB results evaluated using discrete score.

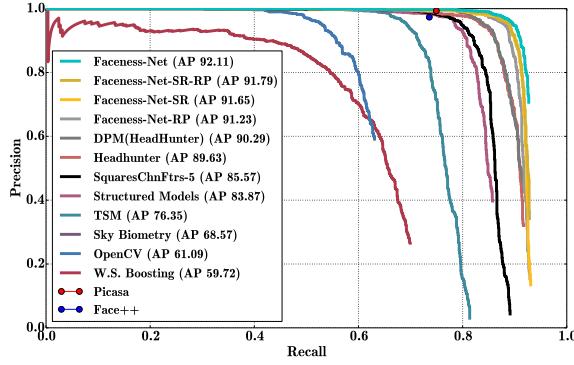


Fig. 13. Precision-recall curves on the PASCAL faces dataset. AP = average precision.

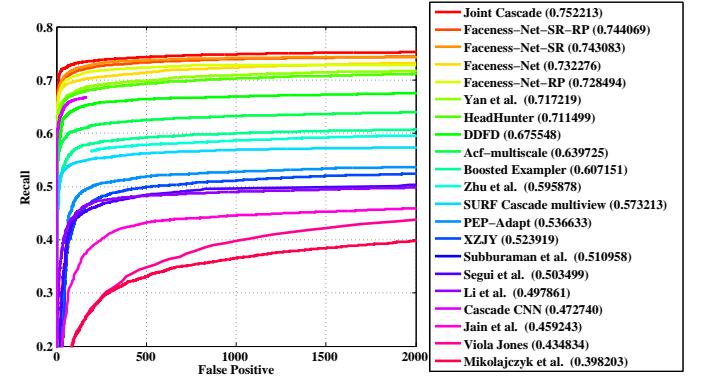


Fig. 15. FDDB results evaluated using continuous score.

[52], [53] on FDDB. For the PASCAL faces and AFW we compare with (1) deformable part based methods, *e.g.* structure model [10] and Tree Parts Model (TSM) [11]; (2) cascade-based methods, *e.g.*, Headhunter [9]. For the WIDER FACE [25] we compare with (1) aggregated channel feature method (ACF) [48]; (2) deformable part based model [9]; (3) cascaded-based method [53].

AFW dataset. Figures 12 shows the precision and recall curves of the compared face detection algorithms on the AFW dataset. We observe that the Faceness-Net and its variants outperform all the previous approaches by a considerable margin. The Faceness-Net-SR and Faceness-Net-RP outperform baseline Faceness-Net showing the effectiveness of sharing representation and region proposal technique. Among all the Faceness-Net variants, Faceness-Net-SP-RP achieves the best performance with a high average precision of 98.05%.

PASCAL faces dataset. Fig. 13 shows the precision and recall curves. The baseline Faceness-Net outperforms its variants and other previous face detection algorithms. Compared with other benchmark datasets, PASCAL faces dataset has fewer number of faces in each image, therefore only requires small number of proposals to achieve a high recall rate. As shown in Fig. 9, the quality of proposals generated by the baseline Faceness-Net is higher than its variants when the number of proposal is less than 200, which leads to its better face detection performance on PASCAL face dataset.

FDDB dataset. The results are shown in Fig. 14 and Fig. 15. Similar with AFW dataset, Faceness-Net and its variants achieve better performance compared with previous algorithms evaluated using the discrete score as shown in the Fig 14. Faceness-Net

baseline achieves 90.99% recall rate, while Faceness-Net-SR and Faceness-Net-RP outperform the baseline Faceness-Net by 0.4% and 0.7%, respectively. Faceness-Net-SR-RP performs best with a large improvement of 1.85% compared with the baseline Faceness-Net. As shown in Fig. 15, the Joint Cascade [5] achieves the best performance evaluated using continues score.

WIDER FACE dataset. WIDER FACE dataset is currently the largest face detection benchmark dataset. We evaluate our algorithm using the external setting because our face detectors are trained without using images in the WIDER FACE dataset. Faceness-Net and its variants yield better performance in all three evaluation settings, namely “Easy”, “Medium”, and “Hard” as shown in Fig. 16. The variants of Faceness-Net outperform baseline Faceness-Net by a considerable margin, suggesting the effectiveness of representation sharing and region proposal techniques.

Discussion: Recent studies [13], [14], [54] achieve better face detection performance on FDDB, AFW, and PASCAL faces datasets compared to our Faceness-Net. The performance gap between Faceness-Net and other methods arises from two aspects, namely, the better modeling of background clutter and stronger supervision signals. Table 4 summarizes the training data and supervision signals used by different algorithms. Faceness-Net is trained on CelebA and AFLW datasets. These datasets are originally proposed for face recognition and facial landmark detection, respectively. The background in CelebA and AFLW is less cluttered and diverse compared with various backgrounds available in WIDER FACE and MS-COCO datasets. In addition, faces in CelebA and AFLW datasets have smaller variations, both

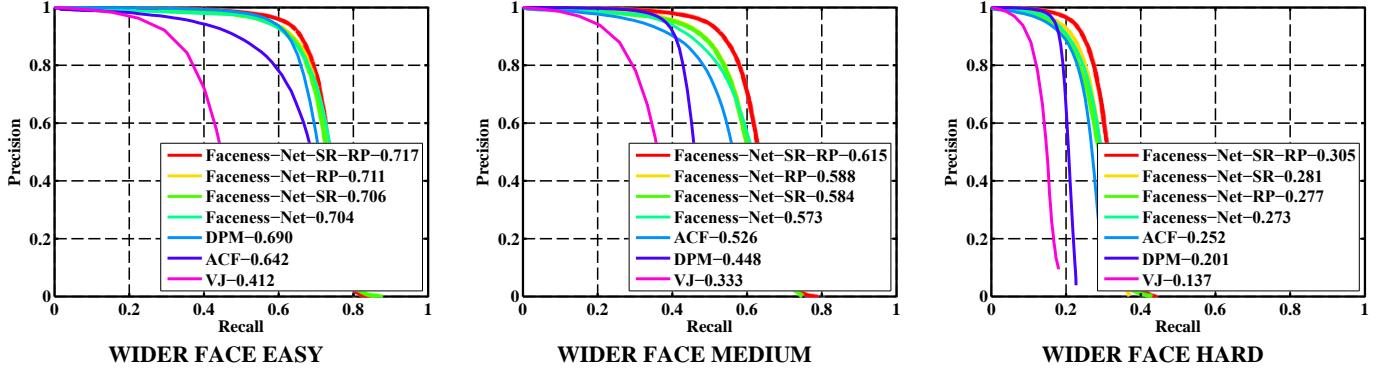


Fig. 16. Precision and recall curves of different subsets of WIDER FACES: Overall Easy/Medium/Hard subsets. AP = average precision.

TABLE 4
A comparison of training data and annotations adopted in state-of-the-art face detection methods.

| | Dataset | #images | #Bounding Boxes | #.landmarks | #Attribute | Clutter Background | ImageNet pretrain |
|------------------|--------------------------|-----------|-----------------|-------------|------------|--------------------|-------------------|
| Faceness-Net | CelebA+AFLW | 180k+13k | 17k | - | 26 | - | - |
| STN [13] | Internal Dataset+MS COCO | 400k+120k | < 400k | 5 | - | ✓ | - |
| Faster-RCNN [14] | WIDER FACE | 13k | 150k | - | - | ✓ | ✓ |
| MTCNN [54] | CelebA+WIDER FACE | 200k+13k | 350k | 5 | - | ✓ | - |

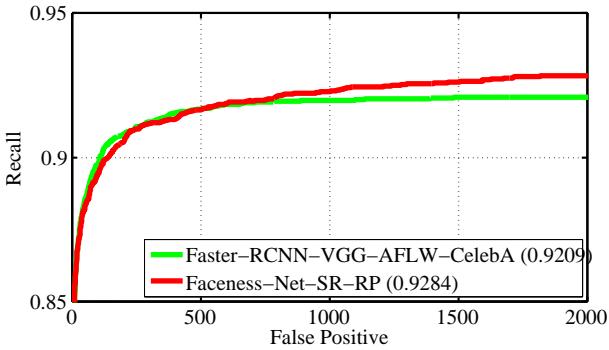


Fig. 17. A comparison of Faster-RCNN face detector [14] and Faceness-Net on FDDB, when both methods adopt the same training data.

in scale and poses, compared to those captured in the WIDER FACE dataset. We use 17k face bounding boxes compared to more than 150k face bounding boxes employed by other methods.

To gain a fairer comparison, we train the Faster-RCNN model presented in [14] using the same training sets (AFLW and CelebA) employed by Faceness-Net. Evaluation is performed on the FDDB dataset. The results are shown in Fig. 17. The Faster-RCNN face detector achieves 92.09% detection rate on the FDDB dataset which is marginally lower than that of Faceness-Net. Note that, Faceness-SR-RP is not finetuned by using ImageNet data, but achieves better performance than Faster-RCNN. This is probably because attribute supervisions are more capable in modeling facial parts.

Apart from using more challenging training images, both STN [13] and MTCNN [54] use facial landmarks to localize face. Facial landmarks indicate explicit location of face parts and thus provide stronger supervisory information than face attributes. Our

method can benefit from these additional factors. Specifically, it is possible to obtain a stronger Faceness-Net detector using facial landmarks based supervision and datasets with more cluttered background.

6 RUNTIME ANALYSIS

The runtime of the proposed Faceness-Net-SR-RP is 40ms on a single GPU⁴. The time includes 10ms to generate faceness proposals with the height of testing image no more than 300 pixels. The efficiency of Faceness-Net-SR-RP is clearly faster than the baseline Faceness-Net since the former shares the layers from Conv1 to Conv4 in its attribute-aware networks. Previous CNN based face detector [16] achieves good runtime efficiency too. 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 4.66% in recall rate (Cascade-CNN 85.67%, our method 90.33%) when the number of false positives is fixed at 167 on the FDDB dataset. The complete recall rate of the proposed Faceness-Net-SR-RP is 92.84% 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. Our Faceness-Net-SR-RP can achieve practical runtime efficiency under the aggressive setting mentioned above, but still with a 0.21% higher recall rate than the Cascade-CNN.

7 CONCLUSION

Different from existing face detection studies, we explored the usefulness of face attributes based supervision for learning a robust face detector. We observed an interesting phenomenon that face part detectors can be obtained from a CNN that is trained on recognizing attributes from uncropped face images, without explicit part supervision. Consequently, we introduced the notion of ‘faceness’ score, which was carefully formulated through considering facial parts responses and the associated spatial arrangements. The faceness score can be employed to re-rank candidate windows of any region proposal techniques to generate a modest set of high-quality face proposals with high recall. With the generated face proposals, we trained a strong face detector that demonstrated promising performance on various face detection benchmark datasets.

4. The runtime is measured on a Nvidia Titan X GPU.

Despite Faceness-Net achieves encouraging face detection performance on benchmark datasets, it may fail to detect tiny faces (with a resolution as low as 20 pixels height). The visual appearance between tiny and normal-size faces exhibits a huge difference. In particular, the facial parts such as eyes, nose or mouth can be barely distinguished from tiny faces, which makes responses produced by attribute-aware networks meaningless. In order to retrieve tiny faces, data augmentation and multi-scale inference may be adopted. Nonetheless, learning scale-invariant representation is still an open problem. In this study, we do not deal with tiny faces explicitly. It is part of our on-going works.

REFERENCES

- [1] R. Vaillant, C. Monrocq, and Y. Le Cun, “Original approach for the localisation of objects in images,” *VISP*, 1994. [1](#) [2](#)
- [2] H. A. Rowley, S. Baluja, and T. Kanade, “Neural network-based face detection,” *TPAMI*, 1998. [1](#) [2](#)
- [3] C. Garcia and M. Delakis, “Convolutional face finder: A neural architecture for fast and robust face detection,” *TPAMI*, 2004. [1](#) [2](#)
- [4] M. Osadchy, Y. Le Cun, and M. L. Miller, “Synergistic face detection and pose estimation with energy-based models,” *JMLR*, 2007. [1](#) [2](#)
- [5] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, “Joint cascade face detection and alignment,” in *ECCV*, 2014. [1](#) [2](#) [9](#) [10](#)
- [6] C. Huang, H. Ai, Y. Li, and S. Lao, “High-performance rotation invariant multiview face detection,” *TPAMI*, 2007. [1](#) [2](#)
- [7] J. Li and Y. Zhang, “Learning surf cascade for fast and accurate object detection,” in *CVPR*, 2013. [1](#) [2](#) [9](#)
- [8] P. Viola and M. J. Jones, “Robust real-time face detection,” *IJCV*, 2004. [1](#) [2](#)
- [9] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, “Face detection without bells and whistles,” in *ECCV*, 2014. [1](#) [2](#) [9](#) [10](#)
- [10] J. Yan, X. Zhang, Z. Lei, and S. Z. Li, “Face detection by structural models,” *IVC*, 2014. [1](#) [2](#) [7](#) [9](#)
- [11] X. Zhu and D. Ramanan, “Face detection, pose estimation, and landmark localization in the wild,” in *CVPR*, 2012. [1](#) [2](#) [7](#) [9](#) [10](#)
- [12] H. Qin, J. Yan, X. Li, and X. Hu, “Joint training of cascaded cnn for face detection,” in *CVPR*, 2016. [1](#) [2](#)
- [13] D. Chen, G. Hua, F. Wen, and J. Sun, “Supervised transformer network for efficient face detection,” in *European Conference on Computer Vision*, 2016. [1](#) [2](#) [10](#) [11](#)
- [14] H. Jiang and E. G. Learned-Miller, “Face detection with the faster R-CNN,” *CoRR*, 2016. [1](#) [9](#) [10](#) [11](#)
- [15] P. Hu and D. Ramanan, “Finding tiny faces,” *arXiv preprint arXiv:1612.04402*, 2016. [1](#)
- [16] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, “A convolutional neural network cascade for face detection,” in *CVPR*, 2015. [1](#) [2](#) [11](#)
- [17] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in *NIPS*, 2015. [1](#) [2](#) [3](#) [5](#) [6](#) [9](#)
- [18] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, “Multi-scale combinatorial grouping,” in *CVPR*, 2014. [2](#)
- [19] V. Jain and E. Learned-Miller, “Fddb: A benchmark for face detection in unconstrained settings,” University of Massachusetts, Amherst, Tech. Rep., 2010. [2](#) [7](#) [9](#)
- [20] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *ICCV*, 2015. [2](#) [7](#)
- [21] 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. [2](#) [7](#)
- [22] S. Yang, P. Luo, C. C. Loy, and X. Tang, “From facial parts responses to face detection: A deep learning approach,” in *ICCV*, 2015. [2](#)
- [23] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, “Selective search for object recognition,” *IJCV*, 2013. [2](#) [3](#) [5](#) [9](#)
- [24] C. Zitnick and P. Dollár, “Edge boxes: Locating object proposals from edges,” in *ECCV*, 2014. [2](#)
- [25] S. Yang, P. Luo, C. C. Loy, and X. Tang, “WIDER FACE: A face detection benchmark,” in *CVPR*, 2016. [2](#) [7](#) [9](#) [10](#)
- [26] M.-H. Yang, D. Kriegman, and N. Ahuja, “Detecting faces in images: A survey,” *TPAMI*, 2002. [2](#)
- [27] Y.-Y. Lin and T.-L. Liu, “Robust face detection with multi-class boosting,” in *CVPR*, 2005. [2](#)
- [28] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh, “Fast object detection with occlusions,” in *ECCV*, 2004. [2](#)
- [29] S. S. Farfade, M. Saberian, and L.-J. Li, “Multi-view face detection using deep convolutional neural networks,” *arXiv*, 2015. [2](#)
- [30] Y. Li, B. Sun, T. Wu, and Y. Wang, “Face detection with end-to-end integration of a convnet and a 3d model,” in *ECCV*, 2016. [2](#)
- [31] M. Opitz, G. Waltner, G. Poier, H. Possegger, and H. Bischof, “Grid loss: Detecting occluded faces,” in *ECCV*, 2016. [2](#)
- [32] P. Arbelaez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, “Multiscale combinatorial grouping,” in *CVPR*, 2014. [2](#) [3](#) [5](#) [9](#)
- [33] C. L. Zitnick and P. Dollár, “Edge boxes: Locating object proposals from edges,” in *ECCV*, 2014. [2](#) [3](#) [5](#) [7](#) [8](#) [9](#)
- [34] B. Alexe, T. Deselaers, and V. Ferrari, “Measuring the objectness of image windows,” *TPAMI*, 2012. [3](#) [5](#)
- [35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *NIPS*, 2012. [4](#)
- [36] M. Zeiler and R. Fergus, “Visualizing and understanding convolutional neural networks,” in *ECCV*, 2014. [4](#)
- [37] 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](#)
- [38] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, “Learning hierarchical features for scene labeling,” *TPAMI*, 2013. [5](#)
- [39] V. Mnih and G. Hinton, “Learning to label aerial images from noisy data,” in *ICML*, 2012. [5](#)
- [40] Y. Sun, Y. Chen, X. Wang, and X. Tang, “Deep learning face representation by joint identification-verification,” in *NIPS*, 2014. [5](#)
- [41] S. Yang, P. Luo, C. C. Loy, K. W. Shum, and X. Tang, “Deep representation learning with target coding,” in *AAAI*, 2015. [5](#)
- [42] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, “Bing: Binarized normed gradients for objectness estimation at 300fps,” in *CVPR*, 2014. [5](#)
- [43] S. Manén, M. Guillaumin, and L. Van Gool, “Prime Object Proposals with Randomized Prim’s Algorithm,” in *ICCV*, 2013. [5](#)
- [44] 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,” University of Massachusetts, Amherst, Tech. Rep., 2007. [7](#)
- [45] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” *IJCV*, 2010. [7](#)
- [46] A. Kae, K. Sohn, H. Lee, and E. Learned-Miller, “Augmenting crfs with boltzmann machine shape priors for image labeling,” in *CVPR*, 2013. [7](#) [8](#)
- [47] S. Zhu, C. Li, C. C. Loy, and X. Tang, “Transferring landmark annotations for cross-dataset face alignment,” *CoRR*, 2014. [7](#)
- [48] B. Yang, J. Yan, Z. Lei, and S. Z. Li, “Aggregate channel features for multi-view face detection,” *CoRR*, 2014. [9](#) [10](#)
- [49] J. Yan, Z. Lei, L. Wen, and S. Li, “The fastest deformable part model for object detection,” in *CVPR*, 2014. [9](#)
- [50] H. Li, Z. Lin, J. Brandt, X. Shen, and G. Hua, “Efficient boosted exemplar-based face detection,” in *CVPR*, 2014. [9](#)
- [51] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang, “Probabilistic elastic part model for unsupervised face detector adaptation,” in *ICCV*, 2013. [9](#)
- [52] X. Shen, Z. Lin, J. Brandt, and Y. Wu, “Detecting and aligning faces by image retrieval,” in *CVPR*, 2013. [9](#)
- [53] V. Jain and E. Learned-Miller, “Online domain adaptation of a pre-trained cascade of classifiers,” in *CVPR*, 2011. [9](#) [10](#)
- [54] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multi-task cascaded convolutional networks,” *CoRR*, 2016. [10](#) [11](#)

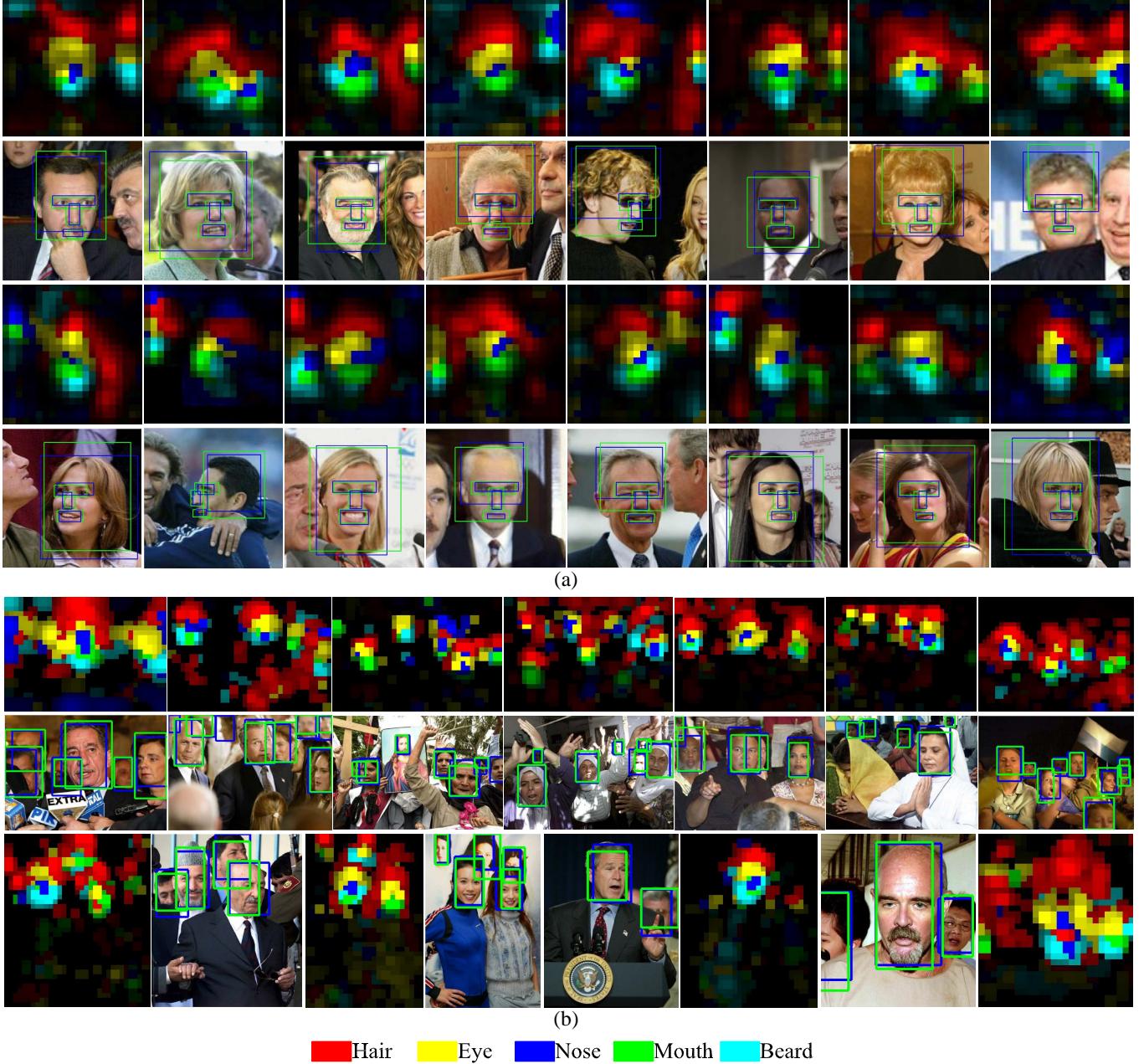


Fig. 18. (a) The first and third row depict the response maps generated by the proposed approach on each part. The second and fourth rows show the part localization results. Ground truth is depicted by the blue bounding boxes, whilst our part proposals are indicated in green. (b) Face detection results on FDDB images. The bounding box in green is detected by our method while ground truth is printed in blue. We show the partness maps as reference. The results shown in (a) and (b) are generated using the baseline Faceness-Net.

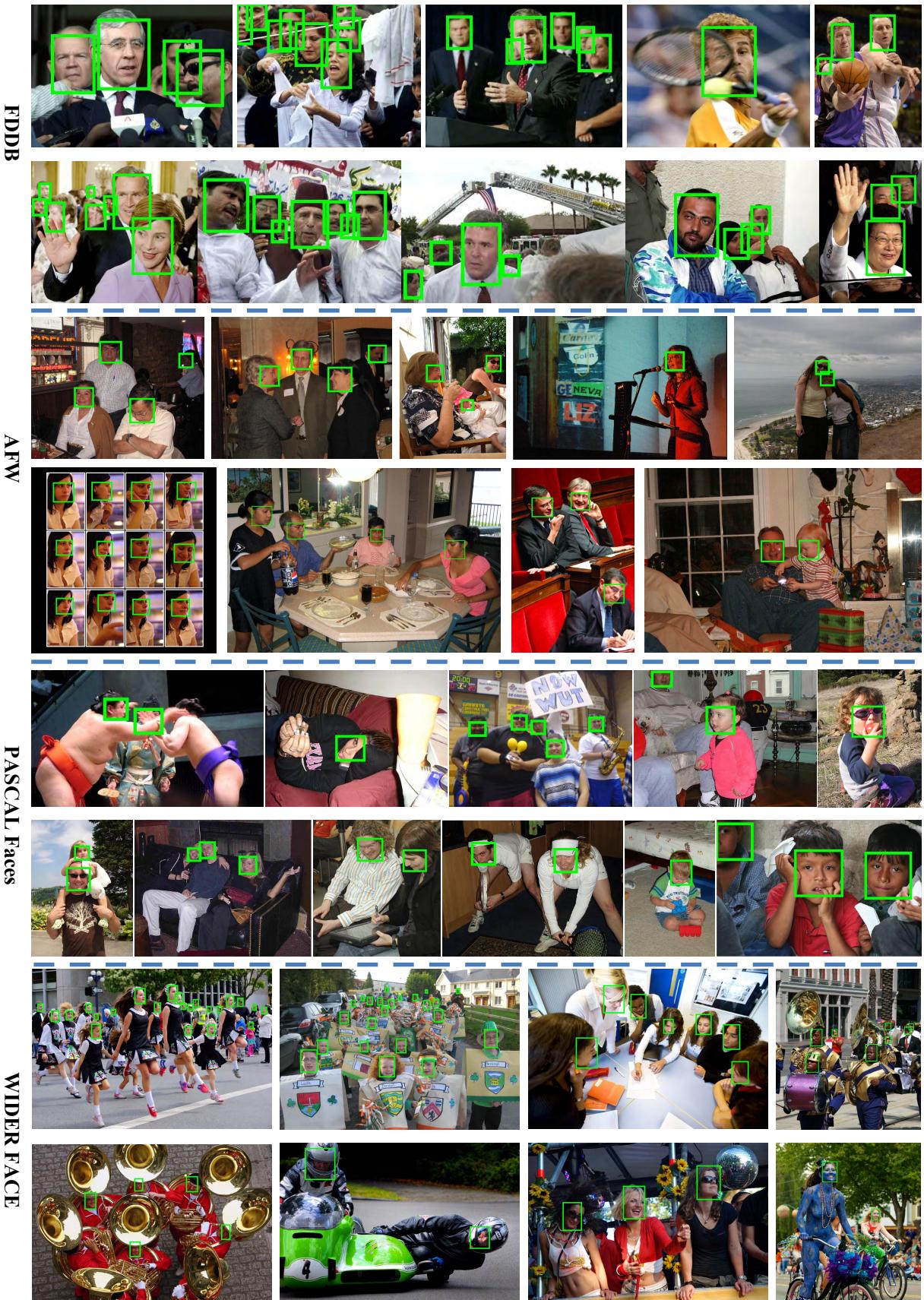


Fig. 19. Qualitative face detection results by Faceness-Net on FDDB (first and second rows), AFW (third and fourth rows), faces (fifth and sixth rows), WIDER FACE (last two rows).