International Journal of Computer Vision (2019) 127:560–578
https://doi.org/10.1007/s11263-019-01157-5

Hierarchical Attention for Part-Aware Face Detection

Shuzhe Wu[1,2] · Meina Kan[1] · Shiguang Shan[1,2,3] · Xilin Chen[1,3]

Received: 15 February 2018 / Accepted: 29 January 2019 / Published online: 2 March 2019

Abstract
Expressive representations for characterizing face appearances are essential for accurate face detection. Due to different poses, scales, illumination, occlusion, etc, face appearances generally exhibit substantial variations, and the contents of each local region (facial part) vary from one face to another. Current detectors, however, particularly those based on convolutional neural networks, apply identical operations (e.g. convolution or pooling) to all local regions on each face for feature aggregation (in a generic sliding-window configuration), and take all local features as equally effective for the detection task. In such methods, not only is each local feature suboptimal due to ignoring region-wise distinctions, but also the overall face representations are semantically inconsistent. To address the issue, we design a hierarchical attention mechanism to allow adaptive exploration of local features. Given a face proposal, *part-specific attention* modeled as learnable Gaussian kernels is proposed to search for proper positions and scales of local regions to extract consistent and informative features of facial parts. Then *face-specific attention* predicted with LSTM is introduced to model relations between the local parts and adjust their contributions to the detection tasks. Such hierarchical attention leads to a part-aware face detector, which forms more expressive and semantically consistent face representations. Extensive experiments are performed on three challenging face detection datasets to demonstrate the effectiveness of our hierarchical attention and make comparisons with state-of-the-art methods.

Keywords Hierarchical attention · Face detection · Object detection · Deformation · Part-aware

1 Introduction

Face detection is a fundamental step for facial information processing, as it has direct influences on subsequent tasks such as face recognition, face anti-spoofing, face editing, face expression analysis, etc. Therefore, an accurate face detector is widely demanded in practical applications. Faces in unconstrained practical scenarios generally exhibit substantial appearance variations due to different poses, scales, illumination, occlusion, etc, and thus make face detection in the wild still a challenging task.

To handle the complicated face variations, most of contemporary face detectors adopt the powerful CNNs, which are highly non-linear models and can learn effective representations of faces automatically from data. The CNN-based face detectors can be roughly categorized into three types according to the generation of face proposals. *The first type* adopts the conventional sliding-window paradigm to enumerate all positions and scales exhaustively, e.g. deep dense face detector (DDFD) (Farfade et al. 2015), Cascade CNN (Li et al. 2015). Compared with conventional methods, they simply switch from hand-crafted features to CNN-learned features. *The second type* densely positions pre-defined anchor boxes[1] of various scales and aspect ratios at different convolutional

Communicated by Xiaoou Tang.

✉ Shiguang Shan
sgshan@ict.ac.cn

Shuzhe Wu
shuzhe.wu@vipl.ict.ac.cn

Meina Kan
kanmeina@ict.ac.cn

Xilin Chen
xlchen@ict.ac.cn

[1] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology (ICT), CAS, Beijing 100190, China

[2] University of Chinese Academy of Sciences (UCAS), Beijing 100049, China

[3] CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai 200031, China

[1] Some papers also call such boxes as "default boxes". Since both default box and anchor box essentially indicate the same thing, hereinafter we use *anchor box* for consistency.

layers as face proposals, e.g. single stage headless face detector (SSH) (Najibi et al. 2017), single shot scale-invariant face detector (S^3FD) (Zhang et al. 2017b). The densely-placed anchor boxes are similar to sliding windows, but for classification, these detectors feed into CNN the whole image instead of each face proposal. They attach convolutional predictors to CNN to classify each anchor box based on their corresponding convolutional features. The detection results are produced by processing the whole image once, featuring a single shot style. *The third type* of CNN-based face detector adopts a particular proposal method to generate a small set of face proposals, which are fed into a (sub-)network for further classification. Jiang and Learned-Miller (2017) propose a face detector based on the Faster R-CNN framework (Ren et al. 2015). It first uses region proposal network (RPN) to generate face proposals, and then compute the representations of fixed dimension for face proposals of variable sizes using region of interest (RoI) pooling, which are taken as input of subsequent layers for classification. Such methods perform detection in two steps, allowing a region to be examined and refined for two times, and therefore usually have advantages in terms of accuracy.

The CNN-based methods have achieved great success in face detection thanks to the powerful representation learning capability of CNNs, but there still exist limitations in the feature computation process. The feature extraction based on CNNs are mainly composition of convolution and pooling operations. The identical convolutional kernels or pooling methods are applied to all face proposals for feature aggregation, which intrinsically makes no distinctions between different proposals and between the local regions within. Moreover, the uniformly computed local features are simply concatenated to form representations of face proposals, which are taken as equally important and directly fed into predictors for classification.

In practical scenarios, faces have complicated variations in appearances, and the details of facial regions differ from each other and vary from one face to another. Some local regions mainly cover non-face areas such as background or other objects, while others contain facial parts, which can appear at varied scales and relative positions, and even in different shapes. Such region-wise distinctions give two hints about feature extraction. First, to obtain effective information from distinct local regions, one should adopt different ways of perception by focusing on proper positions and at consistent scales. This not only allows each local feature to be computed in an optimal way but also helps maintain semantic consistency among representations of different faces. Second, different local regions have divergence in their roles for detection purpose, and should be treated unequally when constructing the face representation. For non-face regions, they can be distractions misleading the detector or support providing the detector with context information. For facial part regions, distinctive parts usually act as strong evidences for distinguishing faces from non-faces, while the rest tend to play a minor role of verifying the decision and adding to the confidence. Therefore, the contributions of different local features to the face detection tasks should be adjusted adaptively according to their contents. Based on these analysis, we can conclude that the uniform feature extraction in current CNN-based methods is suboptimal with respect to both local feature extraction and whole face representation.

There have been face detectors based on part modeling that allow to adaptively extract local features to some extent for different faces. Deformable part model (DPM) (Felzenszwalb et al. 2010; Mathias et al. 2014) exploits distinct part filters for perception of different local regions and learns their geometric configurations with latent SVM. Joint Cascade (Chen et al. 2014) and funnel-structured cascade (FuSt) (Wu et al. 2017) adopt shape-indexed features guided with facial landmarks. These detectors mainly use hand-crafted features, and have lagged behind those CNN-based ones. Deformable CNN (Dai et al. 2017) allows for object deformations by enhancing the RoI pooling operation by learning offsets for each local region. But similar to previous methods, it only takes position search into consideration, lacking the capability of scale adjustment. Moreover, none of the above methods allow to dynamically adjust contributions of different local features to the final face detection task.

To address the above limitations of current methods, we design a hierarchical attention mechanism for adaptive feature aggregation in face detection. The design consists of *part-specific* and *face-specific* attention, forming a hierarchical structure. Specifically, for each local region (part) of a given face proposal, a specific kernel parameterized with Gaussian distribution is predicted. These part-specific kernels adaptively identify the optimal positions, scales and orientations for feature aggregation, and thus can extract more informative local features. On top of the part-specific attention, an LSTM is adopted to model the relations of the extracted local features, which are used to predict an attention map over the entire face proposal, forming face-specific attention. The attention map distinguishes strong and weak features, and adjusts their contributions to the subsequent classification. Figure 1 illustrates the differences of our method from others in representing one face proposal. With such a design, our proposed **p**art-aware face detector with **hi**erarchical attention (*PhiFace*) can effectively handle both region-wise and face-wise distinctions and construct more expressive face representations with better semantic consistency. Experimental results on three challenging face detection data set, i.e. FDDB (Jain and Learned-Miller 2010), WIDER FACE (Yang et al. 2016a) and UFDD (Nada et al. 2018), show that the proposed PhiFace detector brings prominent improvements and achieves promising accuracy.

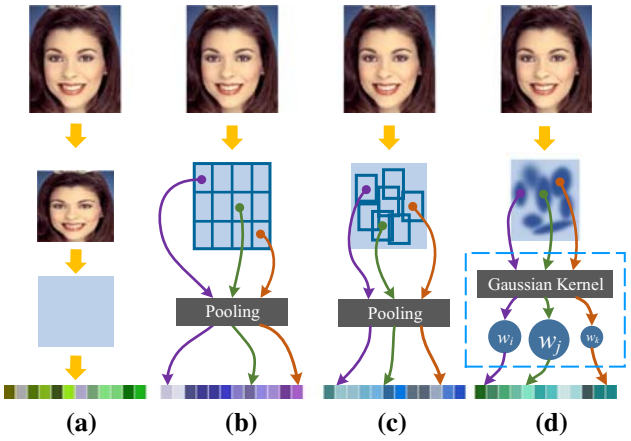


Fig. 1 Comparison of different methods forming the representations of one face proposal. **a** *Resize* the image, e.g. (Li et al. 2015). **b** Pooling with a *regular* grid, e.g. (Ren et al. 2015). **c** Pooling with a *deformable* grid, in which positions of bins can be adjusted, e.g. (Dai et al. 2017). **d** Our *hierarchical attention* using Gaussian kernels with adaptable position, scale and orientation (*part-specific attention*), and attention map that adjusts contributions of different local features of facial parts (*face-specific attention*)

The rest of this paper are organized as follows. Sect. 2 discusses related work on face detection and visual attention. Sect. 3 describes the proposed hierarchical attention mechanism and the designed PhiFace detector in detail. Sect. 4 presents experimental results and analysis. Sect. 5 concludes the paper and discusses the future work.

2 Related Work

Face detection has witnessed steady progress over the past years. The basic framework, feature extractor, classifier and learning scheme have all been significantly improved. Here we briefly review previous researches on face detection in Sect. 2.1 with special emphasis on the face representation. Then in Sect. 2.2, we review works on applying attention to vision tasks, and compare these approaches, though aiming at different tasks, with the proposed hierarchical attention in terms of attention representation and structure.

2.1 Face Detection

From the perspective of face representation, previous face detection methods can be roughly divided into two categories. (1) Methods with rigid templates: These methods extract features within local regions arranged in fixed spatial configuration and usually with identical operations to construct a holistic representation of faces. (2) Methods with part/shape modeling: These methods adaptively handle face deformation by dynamically inferring part features or combining them with holistic features to represent faces.

Methods with rigid templates Since the seminal work of Viola and Jones (2004), face detectors with rigid templates have been extensively explored. The Viola-Jones detector uses Haar-like features selected with AdaBoost to represent faces, which are computationally efficient but are too weak to handle complex face variations. Subsequent works enhance Haar-like features with more complicated patterns (Lienhart and Maydt 2002) and generalize them to more generic linear features (Liu and Shum 2003; Huang et al. 2006). Later, more expressive features such as speeded-up robust features (SURF) (Li and Zhang 2013) and aggregated channel features (ACF) (Yang et al. 2014), which make use of rich gradient and color information, are exploited to improve face detection in unconstrained environment. All of these methods use hand-crafted features, which are computed at pre-defined positions in a fixed way and are used to represent all faces.

With the success of deep learning in vision tasks, the hand-crafted features used in face detection are gradually replaced by ones learned automatically from data with the powerful CNN. The use of CNN in face detection dates back to the work of Vaillant et al. (1994) and Osadchy et al. (2005), which train small CNNs on limited data to distinguish faces from non-faces. More recent CNN-based face detectors adopt much larger networks that are first pre-trained on large scale image classification data and then fine-tuned on face detection data (Zhang and Zhang 2014; Farfade et al. 2015). They achieve comparable performance with those using elaborated and complex hand-crafted features, but yielding a much simpler solution. Li et al. (2015) combine several small CNNs using the conventional cascade structure in a coarse-to-fine manner, leading to an accurate face detector with fast speed. These early works mainly consider to borrow the CNN-learned features but still stick to the conventional face detection pipeline, i.e. the sliding-window paradigm.

As new frameworks emerge in generic object detection, the most recent CNN-based face detectors started to embrace changes in the detection pipeline. Many works adopt the proposal-based Faster R-CNN framework (Ren et al. 2015), e.g. (Jiang and Learned-Miller 2017). To further improve the accuracy on face detection, Wang et al. (2017a) use multi-scale training and online hard example mining (OHEM) (Shrivastava et al. 2016) strategies and add center loss (Wen et al. 2016) as an auxiliary supervision signal for classification. Zhu et al. (2017) integrate multi-scale feature fusion and explicit body contextual reasoning to assist detecting faces in extreme cases such as at very small scale and with heavy occlusion. Chen et al. (2017b) introduce an adversarial mask generator to produce hard occluded face samples to increase the occlusion robustness of the detector. Wang et al. (2017b) adopt the region-based fully convolutional network (R-FCN) (Dai et al. 2016) and use image pyramid to obtain high detection accuracy. Though more advanced detection frameworks with various strategies and new modules are used, these meth-

ods pay little attention to the uniformity issues of feature computation in CNN. The RoI pooling used to construct fixed-dimension representations simply partitions each proposal according to a pre-defined regular grid and pools local features from each grid bin.

There are also face detection researches following the single-shot object detection framework, which does not explicitly generate proposals but uses pre-defined anchor boxes directly. Typical single-shot face detectors include SSH (Najibi et al. 2017), S^3FD (Zhang et al. 2017b), which are inspired from the single-shot multibox detector (SSD) framework (Liu et al. 2016) for generic object detection. To handle scale variations of faces more efficiently, Hao et al. (2017) propose to predict scale histograms to guide the resampling of images, and Liu et al. (2017) design a recurrent scale approximation (RSA) unit to predict feature maps of different scales directly. Such single-shot methods apply convolutional predictors to feature maps to classify anchor boxes. Therefore, each anchor box is simply represented by the convolutional features at the corresponding position, which are uniformly computed with the same kernel across the whole image.

The methods discussed above, either adopting handcrafted features or CNN-learned features, all use rigid templates for classification of faces and non-faces. The face representations are concatenation of local features computed with the same operation at pre-defined positions, regardless of face deformation, resulting in suboptimal modeling of complex face variations.

Methods with part/shape modeling To address the limitations of rigid templates, methods are proposed to integrate part or shape modeling in representation construction, the most typical of which is deformable part model (DPM) (Felzenszwalb et al. 2010). In DPM, an object is considered to be formed by multiple parts, whose positions are inferred online according to the image contents. Such a design has an intrinsic advantage in handling deformation and has been successfully applied to face detection (Mathias et al. 2014). To alleviate the speed issue of DPM, Yan et al. (2014) propose to decompose filters into low rank ones and use look-up table to accelerate computation of histogram of oriented gradients (HOG) features. Similar to DPM, Zhu and Ramanan (2012) design a tree-structured model (TSM) with a shared pool of facial parts, which are defined by facial landmarks, to handle faces in different views, and it jointly learns multiple tasks including face detection, pose estimation and landmark localization. Following this line, Joint Cascade (Chen et al. 2014) and FuSt (Wu et al. 2017) introduce prediction of landmark positions for face proposals besides the classification between faces and non-faces, which are used to extract shape-indexed features to obtain more semantically consistent and thus more discriminative representations for face detection. Although the shape-indexed feature is beneficial for classification between faces and non-faces, it results in loss of localization information needed by the bounding box regression task that is commonly adopted in more recent methods using anchor boxes instead of sliding windows.

Following similar ideas, part or landmark information is also exploited in CNN-based methods to enhance their robustness to diverse face variations. Faceness-Net (Yang et al. 2015) combines five part-specific CNNs for hair, eye, nose, mouth and beard respectively. It improves detection accuracy but incurs heavy computation burden from multiple CNNs. Multi-task cascaded convolutional networks (MT-CNN) (Zhang et al. 2016) adopt a multi-task objective to jointly optimize the classification between faces and non-faces and the landmark localization. Though it obtains improvement from the extra supervision of landmark positions, MT-CNN does not make use of the predicted landmarks for face representation. Chen et al. (2016) design a supervised transformer network (STN) to transform all face proposals into a canonical shape according to facial landmarks. It calibrates different faces so that face variations are reduced from the input. Li et al. (2016) exploit a 3D face model to generate bounding boxes of face proposals, and introduce a configuration pooling operation to extract features according to ten predicted keypoints for subsequent classification, which are similar to the shape-indexed feature but computed with CNN. The above methods all take advantage of extra supervision from part or shape information, which helps reduce face variations at the input or feature level. Under such scheme, it requires sufficient face samples with part or landmark annotations, and it is unclear how many parts are necessary to obtain a good detection accuracy and which set of parts are more effective.

Deformable CNN (Dai et al. 2017) introduces deformable convolution and deformable RoI pooling to handle object deformation. Specifically, for the input of convolution or RoI pooling, it predicts offsets for each element or bin, allowing their positions to be dynamically adjusted according to image contents. The whole model is learned in an end-to-end manner, driven by detection task objectives without extra supervision. Similar to previous methods, deformable CNN focuses on searching positions of parts without explicit mechanism to adjust such as part scales and orientations.

In addition, compared with deformable convolution, which samples a fixed number of positions to cover areas of varying sizes, our method densely samples the input with weights decaying smoothly from the center of Gaussian kernels, leading to three advantages. *First*, our method makes sufficient exploitation of the input. Deformable convolution samples input at dispersed positions, which can be viewed as leaving "holes" in the kernel and thus results in loss of information at those dropped positions. Differently, our Gaussian kernels exploit information at all positions by dense sampling. *Second*, in principle our method is endowed

with better robustness to unexpected noisy or corrupted input values. For deformable convolution, its output could be largely influenced, if it unluckily samples positions that have unexpected values. By contrast, our Gaussian kernels have weights assigned to densely sampled positions, which decay smoothly from the kernel center so that the negative effect of unexpected values could be eased with the smoothing. *Third*, with explicit constraints by the Gaussian density function, our method is able to guarantee consistency among the movements of all sampling positions, which is required to handle various geometric transformations such as rotation. In deformable convolution, however, the movement of each sampling position is independent of each other, making it difficult to guarantee the needed consistency.

Overall, compared to the previous methods with integration of part or shape modeling, the proposed hierarchical attention mechanism not only inherits their advantages, e.g. dynamic inference of part positions, more semantically consistent face representation, driven by detection objectives, end-to-end training, but also features the following new characteristics and capabilities.

- Gaussian distributions are exploited to generate kernels for local feature aggregation, which simulate the human fixation with a smooth decay of attention starting from the center position.
- The kernels are adaptively generated according to contents of local regions with the capability of adjusting both positions and scales and even orientations. Moreover, their receptive fields are adaptive to sizes of proposals.
- Information within a face proposal is sufficiently exploited. The relations of local features are modeled with an LSTM to form an attention map over the entire face proposal, based on which all local features make contributions to the tasks but with different amount of value.

As can be seen, our hierarchical attention takes advantage of both human-vision-like design with prior knowledge, e.g. Gaussian fixations, and data-driven learning, which is endowed with more flexibility and more appropriate way of handling complicated face variations, leading to more expressive face representations for detection. Besides, it is compatible with existing detectors, thus being easily integrated into most of them.

2.2 Visual Attention

When observing an image, one generally pays varied attention to distinct local regions, indicating that distinct regions do not contribute equally to the perception and understanding of the image. The visual attention mechanism has been widely used to associate visual and text contents in tasks like image captioning (Xu et al. 2015; Chen et al. 2017a) and visual question answering (Shih et al. 2016; Yang et al. 2016b; Yu et al. 2017). In these tasks, the attention-based models perform search on the whole input image to identify relevant regions or salient objects, which can be guided by the corresponding text information. By contrast, our hierarchical attention designed for face detection is object-oriented. Specifically, it searches within face proposals to explore informative local facial part features in a finer granularity.

Attention is also widely applied to image and object recognition tasks. Most works use *recurrent models* for attention generation. Ba et al. (2015) exploit recurrent neural network to predict a sequence of glimpses, which are used to localize and recognize multiple digits in the input image. Fu et al. (2017) propose a recurrent attention convolutional neural network (RA-CNN) to progressively attend to more discriminative parts so as to distinguish between fine-grained categories. Wang et al. (2017c) design a recurrent memorized-attention module to iteratively localize object regions to perform multi-label image classification. These methods model attention prediction as a sequential task with the generation of new attention conditioned on the previous one. There are also *non-recurrent* attention models. Hu et al. (2018) propose squeeze-and-excitation networks (SENet) to recalibrate channel-wise feature responses, which can be considered as attention across channels. Zheng et al. (2017) design a multi-attention convolutional neural network (MA-CNN) to localize object parts based on channel grouping. Such works aim at *modeling interdependences and correlations between feature channels*. Ye et al. (2016) design a spatial attention module with rotation and translation transform to reduce variations of hands in viewpoint and articulation for hand pose estimation. Ding et al. (2018) propose to learn attentional face regions for attribute classification under unaligned condition, which is achieved with global average pooling followed by supervision of attribute labels. These works aim to *implicitly align distinct objects*, which is similar to the shape-indexed feature.

In object detection, attention can help construct effective object representations. Hara et al. (2017) propose an attentional visual object detection (AOD) network to predict a sequence of glimpses of varied sizes, forming different views of objects. Li et al. (2017a) introduce a map attention decision (MAD) unit to select appropriate feature channels for objects of different sizes. Li et al. (2017b) propose an attention to context convolution neural network (AC-CNN) to adaptively identify positive contextual information for the object from the global view. He et al. (2017) present a text attention module (TAM) for text detection to suppress background interference. Zhang et al. (2018) exploit channel-wise attention learned with self or external guidances to build occlusion robust representations. Compared with the pro-

posed hierarchical attention, these methods consider from a more global perspective to identify useful contextual information, treating the object as a whole without delving into local regions within objects. Apart from the works mentioned above, there are others that use attention to search for object positions on the whole image, e.g. (Alexe et al. 2012; Caicedo and Lazebnik 2015; Mathe et al. 2016; Jie et al. 2016). Attention in these methods is mainly relevant to generating proposals instead of constructing object representations.

In general, our hierarchical attention mechanism is distinguished from existing visual attention schemes in two aspects. First, a parametric form is adopted to represent attention as Gaussian distributions instead of rectangular boxes and masks. It enjoys good flexibility but with only a few parameters to learn, and directly generates kernels for feature aggregation. Second, the attention is established in a hierarchical structure by combining part-specific attention with face-specific attention. In such a way, the attention over the whole face proposal can be considered as being divided into two simpler parts at different levels, whose search spaces are both smaller and hence could be easier for learning.

3 Hierarchical Attention for Face Detection

This section describes the proposed hierarchical attention mechanism in detail. We adopt the state-of-the-art Faster R-CNN (Ren et al. 2015) as detection framework and design a part-aware face detector with hierarchical attention (PhiFace). The symbols used for describing our method are listed in Table 1 for clarity.

Figure 2 illustrates the schema of our PhiFace detector. Given an input image, first, a backbone CNN is used to compute its feature maps. Then taking the feature maps as input, a convolutional layer followed by two sibling branches for objectness and location prediction, i.e. region proposal network (RPN), is used to generate face proposals. After the proposals are obtained, the proposed hierarchical attention is used to construct expressive representations for each face proposal. Finally, fully-connected layers with the proposal representations as input determine if they are faces or not and refine the locations of all faces, giving accurate detections.

To construct representations of face proposals with our hierarchical attention mechanism, the part-specific and face-specific attention is applied sequentially. The part-specific attention extracts informative local features by searching for optimal ways of feature aggregation (Sect. 3.1), while the face-specific attention adjusts contributions of local features adequately, assigning larger weights to more prominent ones (Sect. 3.2). The former determines *what the local features should be*, while the latter determines *how each local feature should be used*. Overall, they form a two-level hierarchy to construct effective representations of face proposals.

Table 1 Notations used in the definition of the proposed hierarchical attention

Symbol	Meaning
R	Face proposal represented with coordinate of top-left corner and box width and height.
w, h	Width and height of a face proposal.
$r(R)$	Features of face proposal R obtained with RoI pooling or the initial Gaussian kernels.
m, n, T	Pooling width and height for face proposals and the total number of pooling cells, i.e. $T = m \times n$.
$\mu_x, \mu_y, \sigma_x, \sigma_y, \rho$	Means, variances and correlation coefficient of 2D Gaussian distribution.
θ	Parameters of 2D Gaussian distribution/kernel, i.e. $(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$.
θ^0	Initial parameters of 2D Gaussian distribution/kernel. Similar for μ_x^0, etc.
$\Delta\theta$	Change of θ. Similar for $\Delta\mu_x$, etc.
$\mathcal{N}(\cdot)$, $f_\theta(\cdot,\cdot)$	2D Gaussian distribution and its probability density function.
$K_{xy}(\theta)$	Gaussian kernel parameterized with θ.
z_{ij}	Local features obtained with the *part-specific attention*.
g	Global context vector obtained with LSTM.
s_{ij}	Attention map predicted for local features.
u_{ij}	Reweighed features obtained with the *face-specific attention*.
$W_\cdot, b_\cdot$	Weights and biases of neural network layers.
i_t, f_t, o_t, c_t, h_t	Input, forget and output gate, cell activation, hidden vector in LSTM.
$\sigma(\cdot)$	The sigmoid activation function.
$\odot$	Element-wise product between two matrices.
$\otimes$	Inner product between two vectors.

Common symbols like π and symbols used only for substitution, e.g. A in Eq. (2), are omitted for simplicity

3.1 Look into the Local: Part-Specific Attention

The representation of face proposals is essential for face detection. In general, a proposal is represented by features extracted from the local regions within it. Specifically, the proposal is first partitioned into small local regions according to a predefined configuration, and then features within each local region are aggregated to obtain local features of the proposal. Finally, all local features are concatenated to form the representation.

Existing methods process all local regions of a proposal uniformly. Taking the RoI pooling used in the original Faster R-CNN for example, it divides each proposal into multiple rectangular local regions (also called as bins) according to

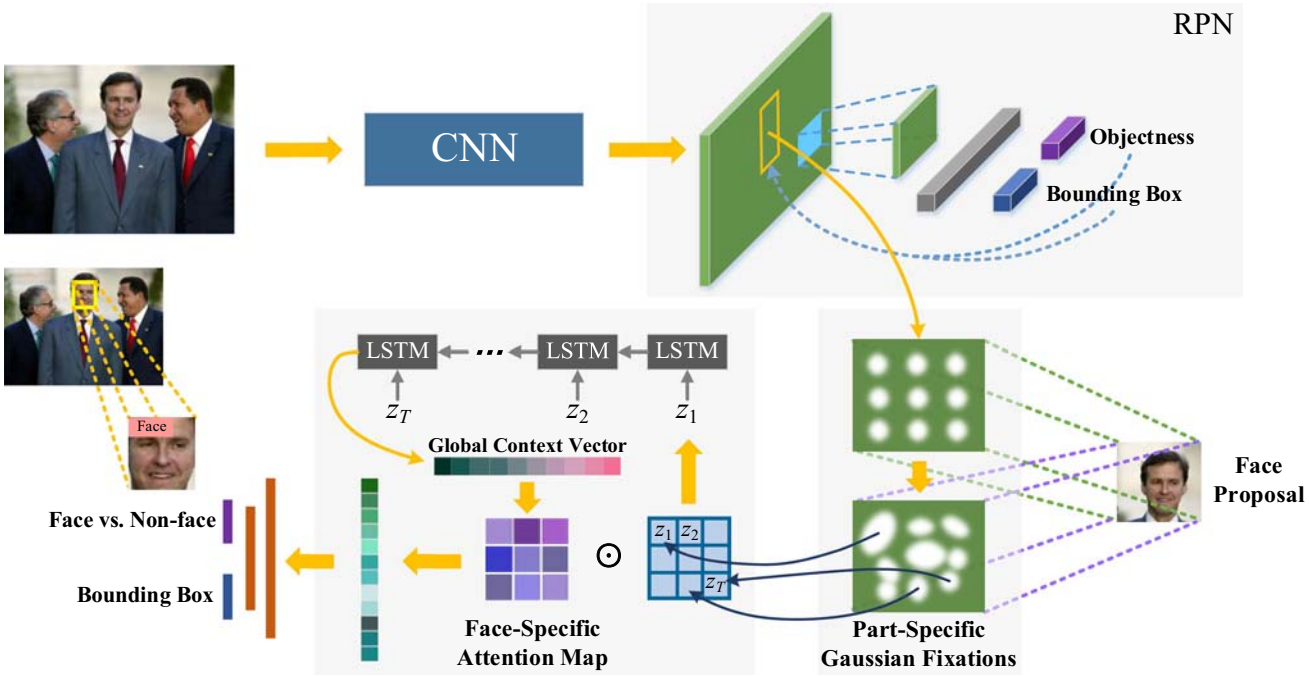


Fig. 2 Schema of the proposed PhiFace detector. Given an input image, a set of face proposals are generated by RPN. For each face proposal, first part-specific attention is applied, and the adaptively predicted Gaussian kernels with varied positions, scales and orientations are used to extract informative local features. Then the local features are sent to an LSTM for global context encoding, based on which face-specific attention maps are generated to adjust contributions of different local features, constructing representations of face proposals. Finally, the representation of each face proposal is fed into subsequent layers for classification (face vs. non-face) and bounding box regression

a fixed $m \times n$ grid. Within each bin, the max pooling is identically exploited for feature aggregation. As a result, all local features are computed at fixed positions and scales with fixed operations, which ignores the diversity of local regions, resulting in sub-optimal representation.

To handle the diversity of local regions within a face proposal, our part-specific attention aims to look into the local to adaptively generate kernels for feature aggregation. Considering that different face proposals are in varying sizes, the kernels for feature aggregation are supposed to be adjustable in sizes, which makes general convolution operation using *fixed-size* kernels not feasible in our case. Moreover, the kernels need to be learnable with gradient descent so that it can learn the rules of adjusting its size according to a given proposal together with the optimization of the rest of the model. To satisfy the adjustability and learnability condition, the kernel is expected to be parameterized with a fixed set of hyperparameters independent of proposal sizes and a differentiable rule for generating its weights and controling its size. Based on the analysis above, we use Gaussian kernels for feature aggregation to implement the part-specific attention (detailed below). For a Gaussian kernel, its position, scale (i.e. size) and orientation are controled by the means, variances and correlation coefficient respectively, endowing the kernel with adjustability. The kernel weights are defined with the Gaussian density function, which is differentiable with respect to the five parameters, allowing them to be optimized using gradient descent.

Specifically, for each local region in a face proposal, simulating human fixations, a kernel parametrized with 2D Gaussian distribution is generated on the fly. Denote the face proposal from RPN as $R \in \mathbb{R}^{w \times h}$ (here the channel dimension is omitted for simplicity), which is divided into $m \times n$ local regions spatially, the parameters of the Gaussian distribution for a local region as $\theta = (\mu_x, \mu_y, \sigma_x, \sigma_y, \rho) \in \mathbb{R}^5$, the corresponding kernel as $K(\theta) \in \mathbb{R}^{w \times h}$, and the aggregated feature as $z \in \mathbb{R}^{m \times n}$. Then the feature aggregation for each local region is formulated as follows:

$$z_{ij} = \sum_{x,y} K_{xy}(\theta_{ij}) \cdot R_{xy}, \tag{1}$$

where i, j are indices of local regions and x, y are indices of positions on the proposal with $1 \le i \le m$, $1 \le j \le n$, $0 \le x < w$, and $0 \le y < h$. Thus $K_{xy}(\theta_{ij})$ indicates the kernel weight at position (x, y) of the (i, j)-th local region parameterized with $\theta_{i,j}$. Note that even though the kernel is defined to be of the same size as the proposal for ease of implementation, it gives small weights to positions distant from the center, thus mainly aggregating features around the corresponding local region.

Definition of Gaussian kernels For each local region, the weights of the kernel for feature aggregation is controled by a 2D Gaussian distribution. Note that Gregor et al. (2015) also use Gaussian distributions for attention but their formulation is different from ours. As for our part-specific attention over local region, denote the Gaussian distribution as $\mathcal{N}(\theta) = \mathcal{N}(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$, and the corresponding probability density function as $f_\theta(x, y)$ defined in the following equations:

$$f_\theta(x, y) = \frac{1}{Z} e^{-\frac{A}{2(1-\rho^2)}}, \tag{2}$$

$$A = \frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x \sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}, \tag{3}$$

$$Z = 2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}. \tag{4}$$

The weight at position (x, y) of kernel K is defined as:

$$K_{xy}(\theta) = f_\theta(x, y), \tag{5}$$

where $0 \le x < w$ and $0 \le y < h$. For the generated kernel, the mean μ_x and μ_y determine the position to focus on. The standard deviation σ_x and σ_y, which control how quickly the weights around the mean position decay to zero, determine the scale for feature aggregation. And the correlation coefficient ρ, which adjusts the shape of the kernel, characterizes the orientation of the focus. With Gaussian distribution, the kernels simulate human fixations with smooth decay of attention starting from the center position.

After the kernels are generated, the weights in it are normalized so that they sum to one. This ensures that different kernels have consistent magnitude of weights.

Part-specific attention With the Gaussian kernels defined above, the part-specific attention is achieved by predicting the parameters of Gaussian distributions on the fly. For each of the $m \times n$ local regions, the Gaussian distribution is initialized to be in a circular shape focusing at the region center. Then the values of change with respect to the initial parameters are predicted, allowing the distribution to adapt its position, scale and shape according to contents of regions. As for the predictor, a fully-connected layer can be used, which takes the aggregated features of the proposal as input. The features can be obtained either by the RoI pooling or with the kernels generated by the initial Gaussian distributions.

Denote the aggregated features as r. The values of change $\Delta\theta$ is computed as follows:

$$\Delta\theta = \tanh(W \cdot r(R) + b), \tag{6}$$

where W and b are the weight matrix and bias vector of the fully-connected layer respectively. The $\tanh(\cdot)$ is used as the activation function to allow both positive and negative values of parameter changes. To avoid illegal parameter values such as a negative standard deviation σ_x, the $\tanh(\cdot)$ output is linearly re-scaled to a suitable positive range, e.g. for $x = \tanh(\cdot) \in [-1, 1]$, it can be rescaled to $[0.1, 0.2]$ via $0.05x + 0.15$. With the predicted $\Delta\theta$, the adapted distribution parameters are obtained as below:

$$\mu_x = \mu_x^0 + \Delta\mu_x \cdot w, \tag{7}$$

$$\mu_y = \mu_y^0 + \Delta\mu_y \cdot h, \tag{8}$$

$$\sigma_x = \sigma_x^0 + \Delta\sigma_x \cdot w, \tag{9}$$

$$\sigma_y = \sigma_y^0 + \Delta\sigma_y \cdot h, \tag{10}$$

$$\rho = \rho^0 + \Delta\rho, \tag{11}$$

where $\theta^0 = (\mu_x^0, \mu_y^0, \sigma_x^0, \sigma_y^0, \rho^0)$ are the initial parameters with location at the region center. Note that one only needs to predict $\Delta\theta$ to generate the Gaussian kernels and the θ^0 are pre-defined constants. During model training, the predictor will learn to identify the optimal attention, i.e. optimal θ, for each local region with guidance from the objective of the face detection task. Hence with this part-specific attention scheme, the contents of each local region are sufficiently explored and appropriately aggregated, producing informative local features. Besides, since the part-specific attention aims at extracting informative features to describe only the local facial characteristics of the face proposal, the search scope of positions and scales can be constrained to be small in practice, i.e. they will not move with very large offsets such as from the leftmost to the rightmost.

For the initial configuration of region division, there are also other choices apart from rectangular grid. For example, one can make the regions equally spaced on concentric circles like in binary robust invariant scalable keypoints (BRISK) (Leutenegger et al. 2011) and fast retina keypoint (FREAK) (Alahi et al. 2012).

Representation with Gaussian Fixations By performing feature aggregation in each local region with the predicted Gaussian kernel defined in Eq. (5), each face proposal can be represented by all the obtained local features of facial parts as follows:

$$z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1n} \\ z_{21} & z_{22} & \cdots & z_{2n} \\ \vdots & & & \vdots \\ z_{m1} & z_{m2} & \cdots & z_{mn} \end{bmatrix}, \tag{12}$$

where z_{ij}, defined in Eq. (1), is the informative local feature of the (i, j)-th local region observed with the Gaussian fixations. As described above, each local feature is obtained by adaptively determining the Gaussian distributions with the optimal location, scale and orientation according to the

region contents, which can better characterize the diversity of the local region and also improve semantic consistency. Therefore, this kind of part-specific attention can achieve informative representation of all local regions within a face proposal.

3.2 View from the Global: Face-Specific Attention

After the local features of facial parts are computed, a face proposal is represented by combining all the extracted features. Current methods generally concatenate all local features directly to form the representation, which assumes all local features are equally effective for the detection tasks. As analyzed in Sect. 1, this is not for sure. On the one hand, the distinctions of region contents will lead to divergence in the role of different regions in detection, e.g. noisy background regions can be misleading, while facial part regions provide evidence for categorization. On the other hand, distinct proposals differ in appearances and may exhibit varied preferences for local regions to perform detection, e.g. one face may be well detected by carefully observing the eyes, while another by focusing on the mouth.

To better characterize face proposals with local features of facial parts, we further introduce a face-specific attention scheme, viewing from the global, to construct more expressive face representations. Specifically, an attention map over the entire proposal is predicted, which assigns an appropriate weight to each local feature. This adjusts the contributions of distinct local features to the tasks adaptively, distinguishing between prominent and defective features. Denote the attention map as $s \in \mathbb{R}^{m \times n}$, whose element s_{ij} is the attention weight of the feature z_{ij}. Then the representation u of the proposal with face-specific attention is obtained as below:

$$u = s \odot z = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & & & \vdots \\ s_{m1} & s_{m2} & \cdots & s_{mn} \end{bmatrix} \odot \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1n} \\ z_{21} & z_{22} & \cdots & z_{2n} \\ \vdots & & & \vdots \\ z_{m1} & z_{m2} & \cdots & z_{mn} \end{bmatrix}, \tag{13}$$

where $\odot$ indicates element-wise product. As can be seen from Eq. (13), each local feature is adaptively weighted to form a comprehensive representation of the proposal.

The correct judgement on the effectiveness of local features requires a comprehensive and overall consideration of all the features. If not appropriately predicted, the attention map is prone to incur little or even negative influences, resulting in degraded face representation. To obtain an effective attention map, the proposed face-specific attention scheme is designed with an encoding process, in which an LSTM (Hochreiter and Schmidhuber 1997) is adopted to model the relations between all local features, summarizing them into a global context vector. The local features are fetched according to the initial configuration from left to right and from top to bottom, and sent to the LSTM sequentially. The LSTM used here is as described in (Zaremba and Sutskever 2014), which does not have peephole connections. The composition functions are defined as:

$$i_t = \sigma(W_{hi} h_{t-1} + W_{zi} z_t + b_i), \tag{14}$$
$$f_t = \sigma(W_{hf} h_{t-1} + W_{zf} z_t + b_f), \tag{15}$$
$$o_t = \sigma(W_{ho} h_{t-1} + W_{zo} z_t + b_o), \tag{16}$$
$$c_t = f_t \otimes c_{t-1} + i_t \otimes \tanh(W_{hc} h_{t-1} + W_{zc} z_t + b_c), \tag{17}$$
$$h_t = o_t \otimes \tanh(c_t). \tag{18}$$

The t stands for timestamp, which corresponds to the sequence index of a local feature. The i, f, o stand for input, forget and output gate respectively. The c and h are cell activation and hidden vectors. The z denotes input vector, i.e. the local features. The W and b denote the weight matrices and bias vectors of linear transformations. The $\sigma(\cdot)$ is the sigmoid function. Denote the number of local features as $T = m \times n$. Then the global context vector summarizing all local features is defined as $g = [c_T; h_T]$, i.e. the concatenation of the final cell activation and hidden vector. Benefited from the memory mechanism in LSTM, the global context vector can be constructed in a progressive way by observing local features sequentially, allowing comprehensive and overall consideration of all the features.

After the global context vector g is obtained with LSTM, one fully-connected layer with sigmoid activation function is used to predict the attention map as defined in Eq. (19), forming face-specific attention:

$$s = \sigma(Wg + b) \tag{19}$$
$$= \sigma(W \cdot [c_T; h_T] + b). \tag{20}$$

The W and b are the weight matrix and bias vector of the fully-connected layer. With such face-specific attention scheme, the contributions of different local features to the face detection tasks are adjusted adaptively, leading to more expressive representation of face proposals.

For clarity, the symbols used above in the definition of our hierarchical attention are listed in Table 1. We further summarize in Algorithm 1 the steps to construct representations of face proposals using our hierarchical attention, including both the part-specific (described in Sect. 3.1) and face-specific attention. The subscript i in some variables are omitted for simplicity.

To acquire the final face detection results, the obtained representations of face proposals are further fed into a sub-network $\mathcal{M}$ for classification, i.e. face vs non-face, and location refinement. The sub-network can be a set of fully-connected or convolutional layers.

Algorithm 1 Hierarchical attention for face proposals

Input: Face proposals $\{R_1, \cdots, R_N\}$, feature maps F
Output: Representations $\{u_1, \cdots, u_N\}$ of face proposals
for $i \leftarrow 1$ to N **do**
  // (1) Part-specific attention
  Obtain initial features r of R_i with F using RoI pooling
  Predict parameter θ with r using Eq. (6) to (11)
  Generate Gaussian kernel K with θ using Eq. (2) to (5)
  Obtain features z of R_i with K using Eq. (1) and (12)
  // (2) Face-specific attention
  Obtain global context vector g with z using LSTM
  Predict attention map s with g using Eq. (19)
  Obtain features u_i of R_i with s using Eq. (13)
end for

In summary, the overall objective of face detection is defined as the following optimization problem:

$$\min_{K,s} \sum_{D} L(\mathcal{M}(u), [c_{\text{gt}}, l_{\text{gt}}]), \tag{21}$$

where D stands for the training data, c_{gt} and l_{gt} are the groundtruth label of classes (face/non-face) and locations, and L indicates the loss function, i.e. softmax loss for classification and smooth L_1 loss (Girshick 2015) for bounding box (location) regression.

3.3 Optimization

The designed PhiFace detector with hierarchical attention can be trained with stochastic gradient descent in an end-to-end manner to solve Eq. (21).

First, the gradient $\frac{\partial L}{\partial u}$ is easily obtained via backpropagation through $\mathcal{M}$ as in general CNNs. Second, the gradients with respect to the face-specific attention map and the local features are obtained according to the chain rule for differentiation as follows:

$$\frac{\partial L}{\partial s} = \frac{\partial L}{\partial u} \cdot \frac{\partial u}{\partial s}, \tag{22}$$

$$\frac{\partial L}{\partial z} = \frac{\partial L}{\partial u} \cdot \frac{\partial u}{\partial z}. \tag{23}$$

Third, the gradients $\frac{\partial L}{\partial s}$ is backpropagated to the LSTM, whose gradient computation has been extensively studied in the literature. So we only derive the gradients with respect to parameters of Gaussian distributions.

By applying the chain rule for differentiation, one can obtain:

$$\frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial z} \cdot \frac{\partial z}{\partial K} \cdot \frac{\partial K}{\partial f_\theta} \cdot \frac{\partial f_\theta}{\partial \theta}. \tag{24}$$

The gradient $\frac{\partial z}{\partial K}$ can be easily derived from Eq. (1). For simplicity, here we ignore the normalization step, whose derivative is similar to that in softmax normalization operation. So the core part is the $\frac{\partial f_\theta}{\partial \theta}$.

Denote the kernel weight at position (x, y) as $p_{xy} = f_\theta(x, y)$. Define A_1, A_2, A_3, A_4 and A_5 as follows:

$$A_1 = \frac{(x-\mu_x)^2}{\sigma_x^2}, \tag{25}$$

$$A_2 = \frac{(y-\mu_y)^2}{\sigma_y^2}, \tag{26}$$

$$A_3 = -\frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}, \tag{27}$$

$$A_4 = -\frac{1}{2(1-\rho^2)}, \tag{28}$$

$$A_5 = A_4 A. \tag{29}$$

Then the derivative $\frac{\partial f_\theta}{\partial \theta}$, i.e. the one of f_θ with respect to $\theta = (\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ at position (x, y) can be computed with the following equations:

$$\frac{\partial f_\theta}{\partial \mu_x} = 2p_{xy} \cdot A_4 \left[\frac{\rho(y-\mu_y)}{\sigma_x\sigma_y} - \frac{x-\mu_x}{\sigma_x^2}\right], \tag{30}$$

$$\frac{\partial f_\theta}{\partial \mu_y} = 2p_{xy} \cdot A_4 \left[\frac{\rho(x-\mu_x)}{\sigma_x\sigma_y} - \frac{y-\mu_y}{\sigma_y^2}\right], \tag{31}$$

$$\frac{\partial f_\theta}{\partial \sigma_x} = -\frac{p_{xy}}{\sigma_x} \cdot [1 + A_4(2A_1 + A_3)], \tag{32}$$

$$\frac{\partial f_\theta}{\partial \sigma_y} = -\frac{p_{xy}}{\sigma_y} \cdot [1 + A_4(2A_2 + A_3)], \tag{33}$$

$$\frac{\partial f_\theta}{\partial \rho} = p_{xy} \cdot A_4 \left[1 - 4\rho A_5 + \frac{(x-\mu_x)(y-\mu_y)}{(1-\rho^2)\sigma_x\sigma_y}\right]. \tag{34}$$

4 Experiments

To demonstrate the effectiveness of the proposed hierarchical mechanism and compare the designed PhiFace detector with other face detection methods, extensive experiments are performed on three challenging face detection data sets, including FDDB (Jain and Learned-Miller 2010), WIDER FACE (Yang et al. 2016a) and UFDD (Nada et al. 2018). In Sect. 4.1, ablation analysis is performed to validate the part-specific and face-specific attention scheme step by step and visualize the predicted attention for intuitive understanding. Then in Sect. 4.2, we dissect the sources of the improvement of our method by examining the classification and localization errors. In Sect. 4.3, supervision from facial landmarks is explored to compare our design with the shape-indexed feature. In Sect. 4.4, our PhiFace detector is compared with other detectors, showing that it achieves state-of-the-art results.

Data set The FDDB data set (Jain and Learned-Miller 2010) contains 2,845 images and 5,171 annotated faces with bound-

ing ellipses in total. The faces are captured in unconstrained environment and exhibit large variations in pose, scale, illumination, occlusion, etc. For the evaluation on the FDDB, we run our PhiFace detector on all the 2,845 images, and use the official tool to compute true positive rates and the number of false positives, which are used to plot ROC curves and compare with other methods, following the standard protocol.

The WIDER FACE data set (Yang et al. 2016a) is a much larger and more challenging set. There are 32,203 images from 62 different events and 393,703 annotated faces in total. All faces are marked with tight bounding boxes, and they have high degree of variability in various aspects. The whole set is divided into three subset according to difficulty, i.e. easy, medium and hard, and the hard subset contains very challenging and even extreme cases. It is split into a training, validation and test set, containing 12,880, 3,226 and 16,097 images respectively, all of which cover images from easy, medium and hard subset. In all experiments, the models are trained on the training set and evaluated on the validation and test set, as most existing works do. We use the official tool to compute precisions and recalls on the validation set and submit our results as instructed to obtain the results on the test set. The Precision-Recall (PR) curves and average precisions (AP) are used for comparison among different methods.

The unconstrained face detection dataset (UFDD) (Nada et al. 2018) is a recently released set focusing on scenarios that are challenging for face detection but receive little attention in other dataset. It involves seven degradations or conditions including rain, snow, haze, lens distortions, blur, illumination variations and distractors. There are a total of 6,425 images with 10,897 face annotations. In Sect. 4.4, we compare our PhiFace detector with other methods on UFDD, following the *external* protocol (Nada et al. 2018). Specifically, we use WIDER FACE dataset (Yang et al. 2016a) as the external training set and the whole UFDD dataset as the test set, reporting APs using the official evaluation tool.

Implementation details Our PhiFace detector is implemented with the `Caffe` framework (Jia et al. 2014).[2,3] In all our experiments, the same network structure as that of VGG-16 (Simonyan and Zisserman 2014) is used, and the LSTM for face-specific attention has 128 hidden cells. As for parameter initialization, we adopt the ImageNet-pretrained model[2] for all networks. Stochastic gradient descent is adopted to train the network for 70k iterations with an initial learning rate of 0.001 that is decreased to 0.0001 after 50k iterations. For more stable convergence, the network is first pre-trained with only the part-specific attention, and then the face-specific attention is added to fine-tune the whole network in an end-to-end manner.

[2] http://caffe.berkeleyvision.org/.

[3] https://github.com/rbgirshick/py-faster-rcnn.

Table 2 Ablation analysis: effectiveness of the part-specific and face-specific attention

Network	Attention		AP		
	Part-specific	Face-specific	Easy	Medium	Hard
ResNet-50	×	×	0.917	0.872	0.668
ResNet-101	×	×	0.914	0.866	0.662
VGG-16	×	×	**0.929**	0.894	0.710
VGG-16	⋆	×	0.928	0.910	0.759
VGG-16	✓	×	0.926	0.910	0.765
VGG-16	×	✓	0.928	0.902	0.743
VGG-16	✓	✓	0.928	**0.914**	**0.781**

Star (⋆): only position search is enabled. Results on WIDER FACE validation set are reported in terms of AP
Best results are shown in bold

4.1 Ablation Analysis of Hierarchical Attention

In this section, we validate the proposed part-specific and face-specific attention scheme step by step on the WIDER FACE validation set.

Baseline For an in-depth analysis, we first train vanilla Faster R-CNN using different network structures, including VGG-16, ResNet-50 and ResNet-101 (He et al. 2016)[4] for evaluation, among which the one using VGG-16 is the direct baseline of our method. The results are given in Table 2 (top-3 rows). As can be seen, on the easy and medium subset, the vanilla Faster R-CNN can achieve good APs, but when it comes to the hard subset, its performance drops severely. Note that both ResNet-50 and ResNet-101 perform worse than VGG-16. The reason may be that the neurons at the last few layers of the deep ResNets have excessively large receptive fields, which are not appropriate to detect the many small faces in WIDER FACE data set.

Effectiveness of hierarchical attention To prove the effectiveness of both part-specific and face-specific attention, we design three models that adopt different attention schemes: (1) only use part-specific attention with position search (similar to the deformable RoI pooling (Dai et al. 2017)), i.e. both scales and orientations remain unchanged; (2) only use part-specific attention (position, scale and orientation are all learnable); (3) use hierarchical attention, i.e. both part-specific and face-specific attention. The results are given in Table 2 (bottom-4 rows). As shown in the table, the two attention schemes both bring obvious improvements over the baseline model, especially on the challenging hard subset. Allowing the adjustment of scale and orientation brings more gains on top of that from position search. And the face-specific attention further increases the performance. This

[4] ImageNet pretrained models of ResNet are obtained from https://github.com/KaimingHe/deep-residual-networks.

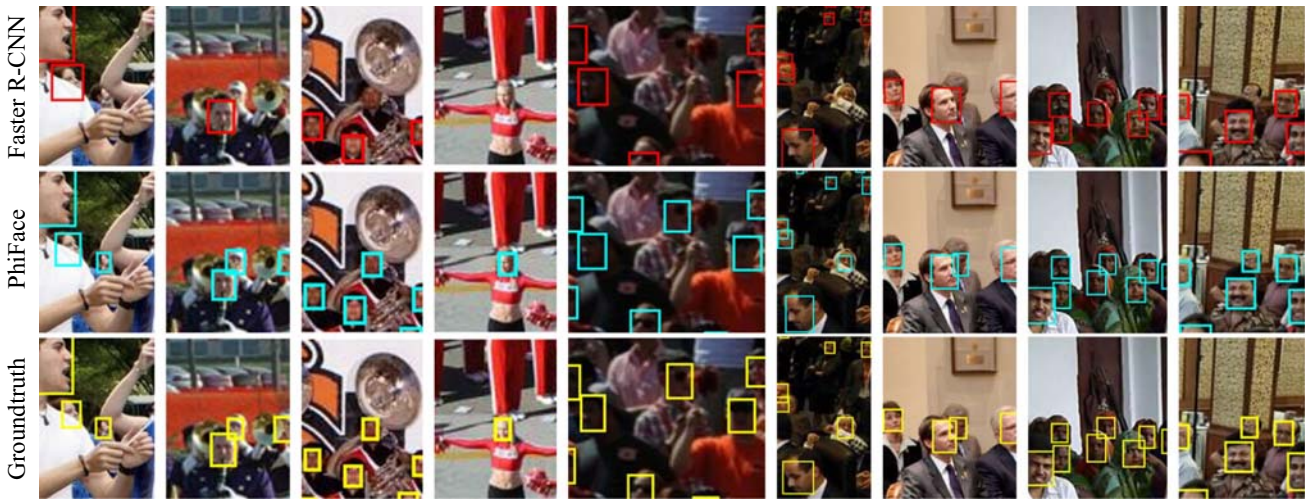


Fig. 3 Examples of the faces that are missed by the Faster R-CNN (row 1) but are recalled by our PhiFace detector (row 2) together with the groundtruth (row 3)

Table 3 Ablation analysis: influence of the initial Gaussian scale

Initial scale (σ_x^0, σ_y^0)	AP		
	Easy	Medium	Hard
0.06	0.926	0.910	0.765
0.10	0.933	0.915	0.765
0.12	0.931	0.914	0.765
0.16	0.931	0.913	0.765

Results on WIDER FACE validation set are reported

Table 4 Ablation analysis: predictor for attention map

Predictor for attention map	AP		
	Easy	Medium	Hard
2FC	0.923	0.894	0.735
Conv + FC	0.926	0.897	0.731
2Conv	0.926	0.898	0.738
Ours	**0.928**	**0.902**	**0.743**

Results on WIDER FACE validation set are reported. Note that 3×3 kernels are used for the convolutional layer (Conv)
Best results are shown in bold

demonstrates that the proposed hierarchical attention mechanism can adaptively handle the complex variations of faces, leading to significant performance improvement. For more intuitive illustration of our improvement, Fig. 3 gives examples of faces that are missed by the Faster R-CNN but are recalled by our PhiFace detector, including faces with variations in size, illumination, occlusion, etc.

Initialization of Gaussian scale As for the part-specific attention, the mean parameters (μ_x, μ_y) of Gaussian distributions can be naturally initialized to be the region center, and the correlation coefficient ρ initialized to zero. But there is no obviously suitable rule to initialize the scale parameters (σ_x, σ_y). Therefore we test different initializations for scale parameters. The results are presented in Table 3. As can be seen, the four models with distinct initial scales achieves very close results. This shows the model can effectively learn to identify the optimal scale, and the part-specific attention is relatively robust with respect to initial scales.

Predictor for attention map As discussed in Sect. 3.2, the face-specific attention map aims to adjust contributions of local features by weighing them from a global perspective, which requires a comprehensive consideration of all the local features and their relations. Therefore, by transforming spatial positions into a sequence, the LSTM is adopted for the face-specific attention to generate a context vector, which enables us to globally model relations between local features. To validate this design, we compare it with other forms of predictors using convolutional (Conv) and fully-connected (FC) layers. For fair comparison, we keep the complexities of different predictors roughly the same by retaining the same dimension of the intermediate output. The results are given in Table 4. As can be seen, our design using LSTM to predict attention map outperforms the others, especially on the Hard subset, demonstrating the effectiveness of its modeling relations among local features.

Comparison with deformable CNN To further validate the effectiveness of the proposed hierarchical attention, we present a comparison with deformable CNN (DCN) (Dai et al. 2017), which can also perform position search for adaptive feature aggregation to handle face variations. Apart from the discussions on DCN in Sect. 2.1, here we experimentally compare DCN with our PhiFace model using VGG-16 and ResNet-50 as the backbone network. We train a Faster

Table 5 Comparison with deformable CNN (DCN)

Network	Model	AP		
		Easy	Medium	Hard
ResNet-50	Baseline	0.917	0.872	0.668
	DCN	0.909	0.876	0.685
	PhiFace	0.926	0.902	0.751
VGG-16	Baseline	**0.929**	0.894	0.710
	DCN	0.927	0.903	0.742
	PhiFace	0.928	**0.914**	**0.781**

Results on WIDER FACE validation set are reported
Best results are shown in bold

R-CNN baseline, DCN and our PhiFace model under the same settings for fair comparison. The results are given in Table 5.[5] As can be seen, with both networks, though DCN achieve obvious improvement over the Faster R-CNN baseline, it still lags behind our PhiFace model. This demonstrates the superiority of the proposed hierarchical attention, which exhibits great flexibility in position, scale and orientation with part-specific Gaussian fixations and further stresses more prominent local features with face-specific attention maps.

Visualization of attention To obtain an intuitive understanding of the proposed hierarchical attention, we present visualization of the predicted attention on different faces in Fig. 4, highlighting local features that make most contributions to the detection tasks. The three rows show respectively attention on: (1) frontal faces, (2) faces with small pose variations, (3) faces with more complex variations. As is shown in the figure, regions around the facial parts, e.g. eyes, nose and mouth, which are crucial for face detection, are automatically identified with our hierarchical attention, thus introducing part-awareness. Moreover, the shape of the Gaussian kernels are adjusted adaptively with different regions on different faces. Like the visual perception of us human, such hierarchical attention mechanism scans the whole image to acquire useful information from the local, and then put more attention on those prominent ones.

Runtime efficiency Though our hierarchical attention, especially the LSTM used in face-specific attention, introduces additional computational cost, it only adds small overheads, since the LSTM has as few as 128 hidden cells. With the input size of 1000 × 600, the average speed (over the 2,845 images in FDDB) of vanilla Faster R-CNN and our PhiFace detector are 116 ms/image and 142 ms/image respectively, i.e. our PhiFace detector only takes extra 26 ms per image.

[5] Results of DCN are obtained with the official code from https://github.com/msracver/Deformable-ConvNets.

Fig. 4 Visualization of attention: local regions that make most contributions to the face detection tasks

4.2 Analysis of Errors

To better understand and analyze the sources of the improvement of our method, we performed a quantitative examination of the classification and localization errors of Faster R-CNN and our method, following the work of Hoiem et al. (2012). Specifically, detections having intersection-over-union (IoU) between 0.1 and 0.5 with groundtruth boxes are considered as localization errors. Other cases, e.g. missed faces due to no matching boxes with IoUs larger than 0.1 and false alarms with IoUs lower than 0.1, are considered as classification errors. The analysis is performed from two perspectives: (1) cause of missed faces, i.e. *why a face box is not recalled*; (2) cause of false alarms, i.e. *why a non-face box is reported as positive*. The ratios of the two types of errors (to all labeled faces and all detections respectively) produced by Faster R-CNN and our PhiFace model are given in Table 6.

As shown in the table, from both perspectives, our PhiFace model achieves a moderate reduction in classification errors but a notable reduction in localization errors. In other words, compared with the Faster R-CNN baseline, our method obtains large improvement in localization quality. This is particularly beneficial for the performance on smaller faces, since the IoUs between the detected and the groundtruth boxes around smaller faces are more sensitive to small shifts in positions and scales, as is also pointed out by Russakovsky et al. (2015) in the context of object detection. Consistently, in previous experiments in Sect. 4.1, our method achieves notable improvement on the Hard subset that contains many small faces.

The probable reasons for our improvement in localization quality on smaller faces lie in 2 folds. *First*, even though the less informative parts of smaller faces do not contain enough

Table 6 Analysis of classification and localization errors of Faster R-CNN and our method on WIDER FACE validation set

Model	Missed faces		False alarms	
	Classification err.	Localization err.	Classification err.	Localization err.
Faster R-CNN	9.36%	6.51%	7.15%	6.65%
PhiFace (ours)	8.99%	5.19%	7.05%	4.57%

Table 7 Comparison between models with and without supervision from facial landmarks

Model	AP_{50}	AP_{55}	AP_{60}	AP_{65}	AP_{70}	AP_{75}	AP_{80}	AP_{85}	AP_{90}	AP_{95}
Baseline	0.994	0.994	0.994	0.993	0.991	0.989	0.963	0.845	0.350	0.010
w/ keypoint	0.993	0.993	0.992	0.992	0.988	0.982	0.950	0.798	0.386	0.023
w/o keypoint	0.995	0.995	0.995	0.994	0.992	0.988	**0.983**	**0.946**	**0.701**	**0.058**

Average precisions (AP) on Helen dataset with different IoU thresholds are reported. AP_x indicates AP computed with IoU threshold of 0.x
Best results with IoU thresholds larger than 0.75 are shown in bold

information for accurately distinguishing between face and non-face boxes, it could still help identify the box boundaries, thus leading to better localization. This is also consistent with our intuition that one can easily mark the face areas even without the rich details of parts. *Second*, compared with the RoI pooling that uses fixed ways of feature aggregation for all face proposals, our hierarchical attention is dynamic and more importantly guided by the localization loss. Therefore, our method can learn to extract better features for localization if needed. Overall, with the guidance of localization loss, our method is able to learn to exploit the less informative parts of smaller faces for better localization quality, thus being able to obtain improvement on the Hard subset that contains many smaller faces.

4.3 Supervision from Facial Landmarks

For the position search in our part-specific attention, one natural idea is to use facial landmarks as supervision. This implies extracting shape-indexed feature as in Joint Cascade (Chen et al. 2014) and FuSt (Wu et al. 2017), which aims to achieve a definite semantic alignment between face boxes. This kind of alignment is beneficial for classification between face and non-face boxes, but it will cause the loss of localization information which is essential for bounding box regression in anchor-based face detectors. For instance, in the case of two distinct boxes overlapping with the same face, they need different calibrating actions to obtain more accurate face boxes. This requirement, however, is difficult to achieve using shape-indexed features. Since the two boxes have almost the same shape-indexed features, the predicted calibrating actions for them will be identical, resulting in incorrect localization of one of them. Different from the shape-indexed feature, our hierarchical attention only performs local position searches without enforcing alignment explicitly. Moreover, as its learning is directly driven by the bounding box regression loss, it is encouraged to extract

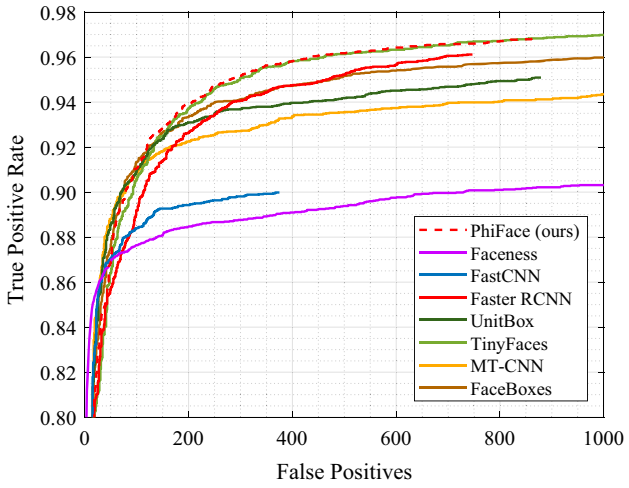


Fig. 5 Comparison between existing methods and ours on FDDB in terms of ROC curves

information that is beneficial not only for classification but also for localization.

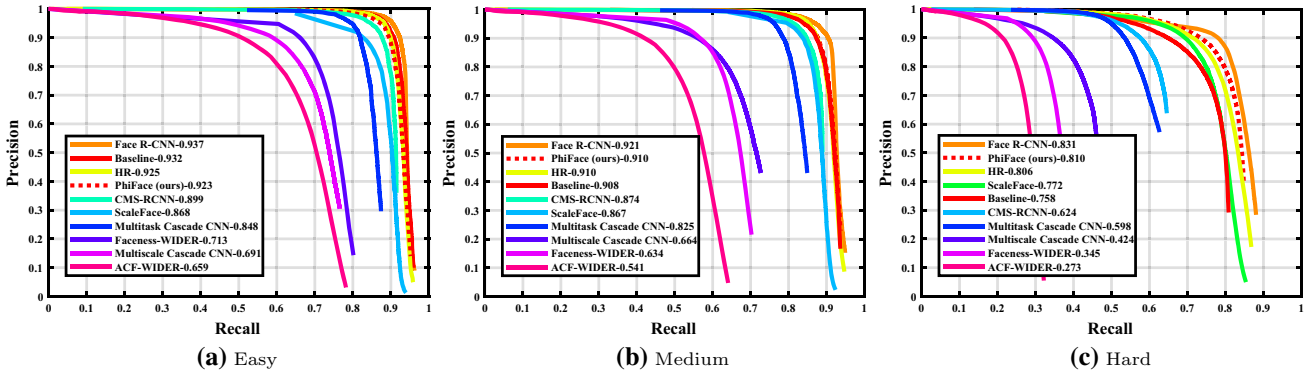
To validate the above arguments and show the advantages of our design over shape-indexed features, we compare models with and without supervision from facial landmarks for part-specific attention. Since face detection datasets like WIDER FACE (Yang et al. 2016a) generally do not contain labels of facial landmarks, we use two face alignment datasets in this experiment. Specifically, we use the Menpo (Zafeiriou et al. 2017) and Helen (Le et al. 2012; Sagonas et al. 2013) dataset with labels of 68 landmarks for training (8,935 image) and testing (2,330 images) respectively. We compare APs of different models using varied IoU thresholds (i.e. imposing different requirements of localization quality). The results are given in Table 7. Note that since the face alignment dataset is relatively easy to perform face detection, the APs are very high particularly with loose IoU thresholds.

As can be seen, the model without supervision from facial landmarks outperforms the other two, showing clear advan-

Table 8 Comparison between existing methods and ours on FDDB data set

Method	Network	Image Pyramid	TPR@FP 100	300	600
Faster RCNN (Jiang and Learned-Miller 2017)	VGG-16	×	89.19	94.06	95.67
UnitBox (Yu et al. 2016)	VGG-16	×	90.97	93.70	94.51
FaceBoxes (Zhang et al. 2017a)	CNN-15	×	**91.34**	94.12	95.40
TinyFaces (Hu and Ramanan 2017)	ResNet-101	✓	90.58	95.07	96.33
PhiFace (ours)	VGG-16	×	91.05	**95.17**	**96.42**

The detection performance is reported as true positive rates (TPR, %) with 100, 300 and 600 false positives (FP)
Best results are shown in bold



(a) Easy **(b)** Medium **(c)** Hard

Fig. 6 Comparison between existing methods and ours on WIDER FACE validation set in terms of PR curves

tages especially with stricter IoU thresholds larger than 0.75. The model using supervision from facial landmarks only obtains comparable APs with those of the Faster R-CNN baseline. These results indicate the importance of localization information and support our analysis above, demonstrating the advantages of our design over shape-indexed features.

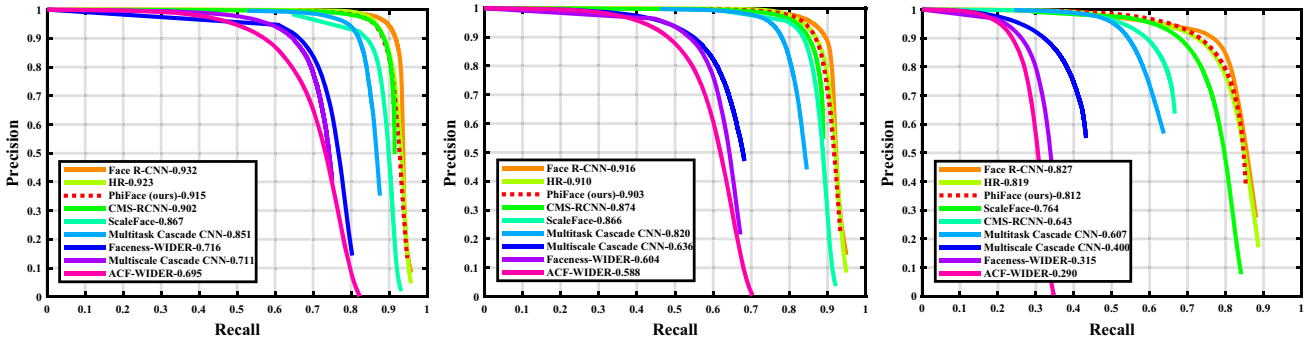
4.4 Comparison with State-of-the-Art

Results on FDDB We validate our PhiFace detector on the FDDB data set (Jain and Learned-Miller 2010), comparing it with other methods[6] including Faceness (Yang et al. 2015), FastCNN (Triantafyllidou and Tefas 2017), Faster R-CNN (Jiang and Learned-Miller 2017), UnitBox (Yu et al. 2016), TinyFaces (Hu and Ramanan 2017), MT-CNN (Zhang et al. 2016), FaceBoxes (Zhang et al. 2017a). The ROC curves are given in Fig. 5, and the true positive rates with 100, 300 and 600 false positives are listed for top-performing methods in Table 8. As can be seen, our PhiFace detector outperforms other methods. Note that although TinyFaces uses the larger ResNet-101 network and exploits image pyramid and multi-scale feature fusion strategy, our PhiFace detector still performs slightly better than it. And compared with the Faster R-CNN, which is the baseline of our method, our PhiFace detector outperforms it by an obvious margin. The superiority of our PhiFace detector comes from the hierarchical attention, which can adaptively handle the complex variations of faces.

Results on WIDER FACE We also compare our PhiFace detector with other methods on the WIDER FACE data set.[7] Since there are a large number of small faces in WIDER FACE, especially in the Hard subset, we remove the `pool4` layer of VGG-16 to obtain a finer feature stride of 8. The methods for comparison include Face R-CNN (Wang et al. 2017a), HR (i.e. TinyFaces) (Hu and Ramanan 2017), CMS-RCNN (Zhu et al. 2017), ScaleFace (Yang et al. 2017), Multitask Cascade CNN (i.e. MT-CNN) (Zhang et al. 2016), Multiscale Cascade CNN (Yang et al. 2016a), Faceness (Yang et al. 2015) and ACF (Yang et al. 2014). Note that here we exclude some methods which adopt image pyramid and flipping during testing for fair comparison, which are drop-in strategies not directly relevant to the method design and are usually not used in Faster R-CNN based detectors. On the validation set, we also report results of baseline Faster R-CNN (Ren et al. 2015) using the adapted VGG-16 network structure.

[6] The results are obtained from the FDDB official website at http://vis-www.cs.umass.edu/fddb/results.html.

[7] The results are obtained from WIDER FACE official website at http://mmlab.ie.cuhk.edu.hk/projects/WIDERFace/WiderFace_Results.html.



(a) Easy **(b)** Medium **(c)** Hard

Fig. 7 Comparison between existing methods and ours on WIDER FACE test set in terms of PR curves

Fig. 8 Examples of detection results on WIDER FACE. The blue rectangles mark faces detected by our PhiFace detector (Color figure online)

The PR curves on WIDER FACE validation and test set are given in Figs. 6, 7 respectively. As can be seen, our PhiFace detector outperforms most methods, obtaining obvious improvement over the baseline, especially on the Hard subset. It achieves comparable performance with that of HR, which feeds three scales of input images into ResNet-101 and adopts multi-scale feature fusion. Face R-CNN shows advantages over other methods, benefiting from the OHEM strategy and auxiliary center loss. As for other methods based on Faster R-CNN, which is the baseline of ours, CMS-RCNN integrates body context reasoning, but our PhiFace detector outperforms it by a large margin, demonstrating the effectiveness of the proposed hierarchical attention. Besides, the various strategies used in these methods are orthogonal to our work, and should be also applicable to the proposed PhiFace detector.

For a more intuitive presentation of detection performance, examples of detection results produced by our PhiFace detector are given in Fig. 8. As shown in the fig-

Table 9 Comparison between existing methods and ours on UFDD dataset

Method	AP
Faster R-CNN (Ren et al. 2015)	0.521
SSH (Najibi et al. 2017)	0.695
S^3FD (Zhang et al. 2017b)	0.725
HR (Hu and Ramanan 2017)	0.742
PhiFace (ours)	**0.746**

APs are reported following the *external* protocol
Best result is shown in bold

ure, our PhiFace detector can well detect faces with different poses, scales, illumination, occlusion, facial expressions, etc. *Results on UFDD* For a further comparison, we evaluate our PhiFace detector on the latest UFDD dataset (Nada et al. 2018) that focusing on many new challenging scenarios. The methods being compared[8] include HR (i.e. TinyFaces) (Hu and Ramanan 2017), SSH (Najibi et al. 2017), S^3FD (Zhang et al. 2017b) and Faster R-CNN (Ren et al. 2015). We use 1x size of original images as input for a single scale testing and the results are given in Table 9. As can be seen, our PhiFace detector outperforms all other methods, demonstrating the effectiveness of the propose hierarchical attention.

5 Conclusions and Future Work

This paper proposes a hierarchical attention mechanism to build expressive face representations for face detection. It consists of part-specific and face-specific attention, forming a hierarchical structure. The part-specific attention with Gaussian kernels simulates human fixations and extract informative and semantically consistent local features of facial parts. The face-specific attention models relations between local features and adjusts their contributions to the face detection tasks. Extensive experiments are performed on the challenging FDDB, WIDER FACE and UFDD data set, and the results show that our PhiFace detector achieves promising performance with large improvement over Faster R-CNN, demonstrating the effectiveness of the proposed hierarchical attention mechanism. For future work, it is an interesting topic to extend and apply our hierarchical attention to generic object detection tasks.

Acknowledgements This research was supported in part by the National Key R&D Program of China (No. 2017YFA0700800), Natural Science Foundation of China (Nos. 61390511, 61650202, 61772496 and 61402443).

[8] The results are obtained from UFDD official website at https://ufdd.info.

References

Alahi, A., Ortiz, R., & Vandergheynst, P. (2012). FREAK: Fast retina keypoint. In *The IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 510–517.

Alexe, B., Heess, N., Teh, Y. W., & Ferrari, V. (2012). Searching for objects driven by context. In *Advances in neural information processing systems (NIPS)*, pp. 881–889.

Ba, J. L., Mnih, V., & Kavukcuoglu, K. (2015). Multiple object recognition with visual attention. In *International conference on learning representations (ICLR)*.

Caicedo, J. C., & Lazebnik, S. (2015). Active object localization with deep reinforcement learning. In *The IEEE international conference on computer vision (ICCV)*.

Chen, D., Ren, S., Wei, Y., Cao, X., & Sun, J. (2014). Joint cascade face detection and alignment. In *European conference on compute vision (ECCV)*, pp. 109–122.

Chen, D., Hua, G., Wen, F., & Sun, J. (2016). Supervised transformer network for efficient face detection. In *European conference on compute vision (ECCV)*, pp. 122–138.

Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T. S. (2017a). SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Chen, Y., Song, L., & He, R. (2017b). Masquer hunter: Adversarial occlusion-aware face detection. arXiv:1709.05188

Dai, J., Li, Y., He, K., & Sun, J. (2016). R-FCN: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems (NIPS)*, pp. 379–387.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. (2017). Deformable convolutional networks. In *The IEEE international conference on computer vision (ICCV)*.

Ding, H., Zhou, H., Zhou, S. K., & Chellappa, R. (2018). A deep cascade network for unaligned face attribute classification. In *The thirty-second AAAI conference on artificial intelligence (AAAI-18)*.

Farfade, S. S., Saberian, M., & Li, L. J. (2015). Multi-view face detection using deep convolutional neural networks. In *International conference on multimedia retrieval (ICMR)*.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *32*(9), 1627–1645.

Fu, J., Zheng, H., & Mei, T (2017) Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Girshick, R. (2015). Fast R-CNN. In *The IEEE international conference on computer vision (ICCV)*.

Gregor, K., Danihelka, I., Graves, A., Rezende, D., & Wierstra, D. (2015). Draw: A recurrent neural network for image generation. *International Conference on Machine Learning (ICML)*, *37*, 1462–1471.

Hao, Z., Liu, Y., Qin, H., Yan, J., Li, X., Hu, X. (2017). Scale-aware face detection. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Hara, K., Liu, M. Y., Tuzel, O., Farahmand, A. M. (2017). Attentional network for visual object detection. CoRR. arXiv:1702.01478

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *The IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 770–778.

He, P., Huang, W., He, T., Zhu, Q., Qiao, Y., & Li, X. (2017). Single shot text detector with regional attention. In *The IEEE international conference on computer vision (ICCV)*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hoiem, D., Chodpathumwan, Y., & Dai, Q. (2012). Diagnosing error in object detectors. In *European conference on compute vision (ECCV)*.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Hu, P., & Ramanan, D. (2017). Finding tiny faces. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Huang, C., Ai, H., Li, Y., & Lao, S. (2006). Learning sparse features in granular space for multi-view face detection. In *The IEEE international conference on automatic face gesture recognition (FG)*, pp. 401–406.

Jain, V., Learned-Miller, E. (2010). FDDB: A benchmark for face detection in unconstrained settings. Technical report UM-CS-2010-009, University of Massachusetts, Amherst.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *ACM international conference on multimedia (MM)*, pp. 675–678.

Jiang, H., & Learned-Miller, E. (2017). Face detection with the Faster R-CNN. In *The IEEE international conference on automatic face gesture recognition (FG)*, pp. 650–657.

Jie, Z., Liang, X., Feng, J., Jin, X., Lu, W., & Yan, S. (2016). Tree-structured reinforcement learning for sequential object localization. In *Advances in neural information processing systems (NIPS)*, pp. 127–135.

Le, V., Brandt, J., Lin, Z., Bourdev, L., & Huang, T. S. (2012). Interactive facial feature localization. In *European conference on compute vision (ECCV)*, pp. 679–692.

Leutenegger, S., Chli, M., & Siegwart, R. Y. (2011). BRISK: Binary robust invariant scalable keypoints. In *The IEEE international conference on computer vision (ICCV)*, pp. 2548–2555.

Li, H., Lin, Z., Shen, X., Brandt, J., & Hua, G. (2015). A convolutional neural network cascade for face detection. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Li, H., Liu, Y., Ouyang, W., & Wang, X. (2017a). Zoom out-and-in network with map attention decision for region proposal and object detection. CoRR. arXiv:1709.04347

Li, J., & Zhang, Y. (2013). Learning SURF cascade for fast and accurate object detection. In *The IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 3468–3475.

Li, J., Wei, Y., Liang, X., Dong, J., Xu, T., Feng, J., et al. (2017b). Attentive contexts for object detection. *IEEE Transactions on Multimedia (TMM)*, *19*(5), 944–954.

Li, Y., Sun, B., Wu, T., & Wang, Y. (2016). Face detection with end-to-end integration of a convnet and a 3D model. In *European conference on compute vision (ECCV)*, pp. 420–436.

Lienhart, R., & Maydt, J. (2002). An extended set of haar-like features for rapid object detection. *International Conference on Image Processing (ICIP)*, *1*, 900–903.

Liu, C., & Shum, H. Y. (2003). Kullback-leibler boosting. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 587–594.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. In *European conference on compute vision (ECCV)*, pp. 21–37.

Liu, Y., Li, H., Yan, J., Wei, F., Wang, X., & Tang, X. (2017). Recurrent scale approximation for object detection in CNN. In *The IEEE international conference on computer vision (ICCV)*.

Mathe, S., Pirinen, A., & Sminchisescu, C. (2016). Reinforcement learning for visual object detection. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Mathias, M., Benenson, R., Pedersoli, M., Van Gool, L. (2014), Face detection without bells and whistles. In *European conference on compute vision (ECCV)*, pp. 720–735.

Nada, H., Sindagi, V., Zhang, H., & Patel, V. M. (2018). Pushing the limits of unconstrained face detection: A challenge dataset and baseline results. CoRR. arXiv:1804.10275

Najibi, M., Samangouei, P., Chellappa, R., & Davis, L. S. (2017). SSH: Single stage headless face detector. In *The IEEE international conference on computer vision (ICCV)*.

Osadchy, M., Miller, M. L., & Cun, Y. L. (2005). Synergistic face detection and pose estimation with energy-based models. In *Advances in neural information processing systems*, pp. 1017–1024.

Osadchy, M., Miller, M. L., & Cun, Y. L. (2005). Synergistic face detection and pose estimation with energy-based models. In *Advances in neural information processing systems*, pp. 1017–1024.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013). A semi-automatic methodology for facial landmark annotation. In *The IEEE conference on computer vision and pattern recognition (CVPR) workshops*.

Shih, K. J., Singh, S., & Hoiem, D. (2016). Where to look: Focus regions for visual question answering. In *The IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 4613–4621.

Shrivastava, A., Gupta, A., & Girshick, R. (2016). Training region-based object detectors with online hard example mining. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. CoRR. arXiv:1409.1556

Triantafyllidou, D., & Tefas, A. (2017). A fast deep convolutional neural network for face detection in big visual data. In *INNS conference on big data*, pp. 61–70.

Vaillant, R., Monrocq, C., & Cun, Y. L. (1994). Original approach for the localisation of objects in images (ip-vis). *IEE Proceedings - Vision, Image and Signal Processing*, *141*(4), 245–250.

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision (IJCV)*, *57*(2), 137–154.

Wang, H., Li, Z., Ji, X., & Wang, Y. (2017a). Face R-CNN. CoRR. arXiv:1706.01061

Wang, Y., Ji, X., Zhou, Z., Wang, H., & Li, Z. (2017b). Detecting faces using region-based fully convolutional networks. CoRR. arXiv:1709.05256

Wang, Z., Chen, T., Li, G., Xu, R., & Lin, L. (2017c). Multi-label image recognition by recurrently discovering attentional regions. In *The IEEE international conference on computer vision (ICCV)*.

Wen, Y., Zhang, K., Li, Z., & Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *European conference on compute vision (ECCV)*, pp. 499–515.

Wu, S., Kan, M., He, Z., Shan, S., & Chen, X. (2017). Funnel-structured cascade for multi-view face detection with alignment-awareness. *Neurocomputing*, *221*, 138–145.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning (ICML)*, pp. 2048–2057.

Yan, J., Lei, Z., Wen, L., & Li, S. Z. (2014). The fastest deformable part model for object detection. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2497–2504.

Yang, B., Yan, J., Lei, Z., & Li, S. Z. (2014). Aggregate channel features for multi-view face detection. In *The IEEE international joint conference on biometrics (IJCB)*, pp. 1–8.

Yang, S., Luo, P., Loy, C. C., & Tang, X. (2015). From facial parts responses to face detection: A deep learning approach. In *The IEEE international conference on computer vision (ICCV)*.

Yang, S., Luo, P., Loy, C. C., & Tang, X. (2016a). WIDER FACE: A face detection benchmark. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Yang, S., Xiong, Y., Loy, C. C., & Tang, X. (2017). Face detection through scale-friendly deep convolutional networks. CoRR. arXiv:1706.02863

Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016b). Stacked attention networks for image question answering. In *The IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 21–29.

Ye, Q., Yuan, S., & Kim, T. K. (2016). Spatial attention deep net with partial pso for hierarchical hybrid hand pose estimation. In *European conference on compute vision (ECCV)*, pp. 346–361.

Yu, D., Fu, J., Mei, T., & Rui, Y. (2017). Multi-level attention networks for visual question answering. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Yu, J., Jiang, Y., Wang, Z., Cao, Z., & Huang, T. (2016). UnitBox: An advanced object detection network. In *ACM on multimedia conference (MM)*, pp. 516–520.

Zafeiriou, S., Trigeorgis, G., Chrysos, G., Deng, J., & Shen, J. (2017). The Menpo facial landmark localisation challenge: A step towards the solution. In *The IEEE conference on computer vision and pattern recognition (CVPR) workshops*.

Zaremba, W., & Sutskever, I. (2014). Learning to execute. CoRR. arXiv:1410.4615

Zhang, C., Zhang, Z. (2014). Improving multiview face detection with multi-task deep convolutional neural networks. In *The IEEE winter conference on applications of computer vision (WACV)*, pp. 1036–1041.

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters (LSP)*, *23*(10), 1499–1503.

Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., & Li, S. Z. (2017a). FaceBoxes: A cpu real-time face detector with high accuracy. In *The IEEE/IAPR international joint conference on biometrics (IJCB)*.

Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., & Li, S. Z. (2017b) S^3FD: Single shot scale-invariant face detector. In *The IEEE international conference on computer vision (ICCV)*.

Zhang, S., Yang, J., & Schiele, B. (2018). Occluded pedestrian detection through guided attention in cnns. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Zheng, H., Fu, J., Mei, T., & Luo, J. (2017). Learning multi-attention convolutional neural network for fine-grained image recognition. In *The IEEE international conference on computer vision (ICCV)*.

Zhu, C., Zheng, Y., Luu, K., & Savvides, M. (2017). CMS-RCNN: Contextual multi-scale region-based CNN for unconstrained face detection. In B. Bhanu & A. Kumar (eds.), *Deep learning for biometrics* (pp. 57–79). Cham: Springer.

Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2879–2886.