Progressively Refined Face Detection Through Semantics-Enriched Representation Learning

Zhihang Li, *Student Member, IEEE,* Xu Tang, Xiang Wu, Jingtuo Liu, and Ran He, *Senior Member, IEEE*

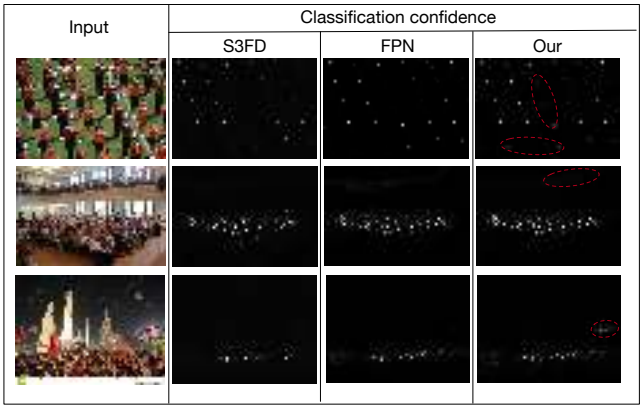
***Abstract*—Feature pyramids aim to learn multi-scale representations for detecting faces over various scales. However, they often lack adequate context over different scales, especially when there are many tiny faces in the wild. In this paper, we propose an attention-guided semantically enriched feature aggregation framework to learn a feature pyramid with rich semantics at all scales for face detection. Specifically, high-level abstract features are directly integrated into low-level representations by skip connections to retain as much semantic as possible. In addition, an attention mechanism is employed as a gate to emphasize relevant features and suppress useless features during feature fusion. Inspired by human visual perception of tiny faces [1], we specially design a deep progressive refined loss (DPRL) to effectively facilitate feature learning. According to the above principles, we design and investigate various feature pyramid frameworks through extensive experiments. Finally, two typical structures named Centralized Attention Feature (CAF) and Distributed Attention Feature (DAF) are proposed for face detection, which are in-place and end-to-end trainable. Extensive experiments across different aggregation architectures on four challenging face detection benchmarks demonstrate the superiority of our framework over state-of-the-art methods.**

***Index Terms*—face detection, object detection.**

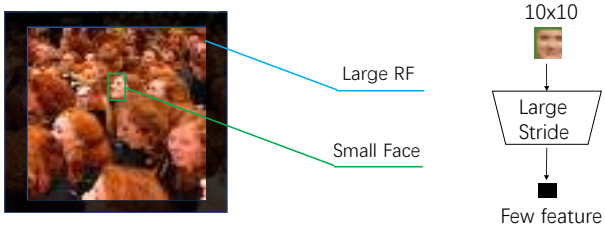
I. Introduction

FACE detection aims to locate faces in a visual scene, which is a fundamental step for subsequent face-relevant tasks, including face tracking [2], face alignment [3], [4] and face recognition [5], [6]. Although impressive progress has been made for decades, face detection in the wild continues to experience various challenges, such as low-resolution imaging, tiny scale faces, large pose variations and occlusions in video surveillance.

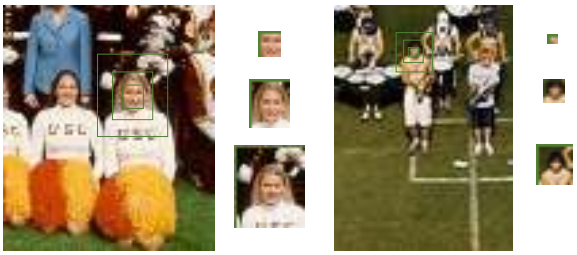
Recently, with the breakthrough of convolution neural network (CNN) in image classification [7] and object detection [8], the performance of face detectors has been substantially improved. As a special case of generic object detection, state-of-the-art face detection algorithms can be roughly divided into two groups: two-stage face detector (Faster RCNN-based methods) [9] and one-stage face detector (SSD-based methods) [10]. Faster RCNN generates proposals by RPN in the first stage and then feeds them into the second stage with ROI pooling for refinement. Although these methods reach a high recall and achieve remarkable results, the training and inference are too time-consuming to be applied in practice. To find a trade-off between speed and precision, SSD exploits inherently multi-scale features from different levels of convolutional outputs in the backbone. With the supervisions



(a) Semantics diffusion



(b) Too large RF for small face



(c) Different scales of extended face regions

Fig. 1. (a) Classification confidence maps of S^3FD, FPN and our method on $conv3\_3$. The brighter the dots, the higher the confidence level. The red circle highlights the differences. (b) Large RF containing more context is not always beneficial to face detection. (c) Visualization of face with different scales of context. A larger region including head and shoulder is easily identified.

Z. Li is with the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: zhihang.li@nlpr.ia.ac.cn)

R. He and X. Wu are with the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: rhe@nlpr.ia.ac.cn; alfredxiangwu@gmail.com). (Corresponding author: Ran He).

X. Tang and J. Liu are with Baidu Inc. (e-mail: tangxu02@baidu.com, liujingtuo@baidu.com)

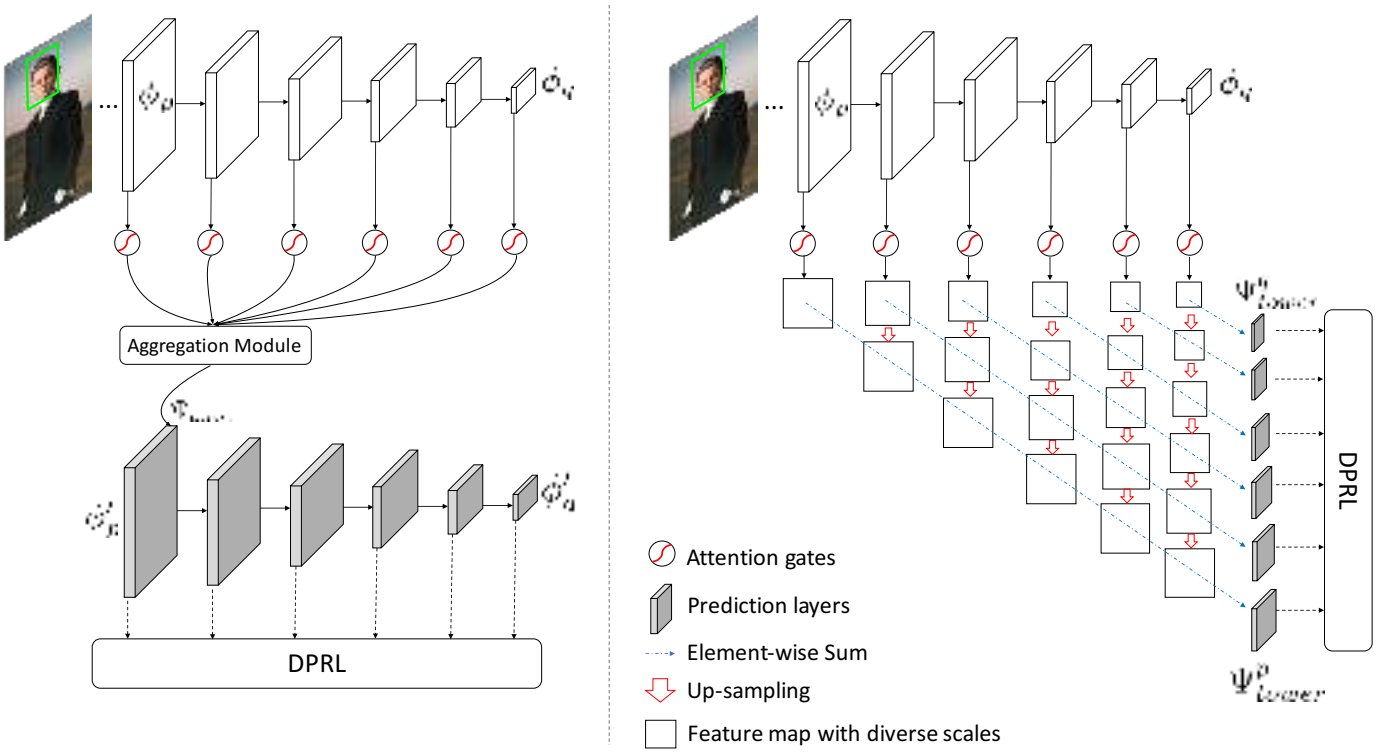
Fig. 2. Example of face detection with the proposed method. In the above image, our method can find 845 faces out of 1000 facial images present. The detection confidence scores are also given by the color bar as shown on the right. Best view in color.

of classification and bounding box regression on multi-scale features, SSD achieves promising performance in real time on GPUs. Because face detection has a high demand for speed in real applications, the one-stage face detector attracts increasing attention [11], [12], [13].

Because Zhang et al[12] demonstrated that the stride size of the lowest anchor-associated layer in SSD is too large (8 pixels in $conv4\_3$), small faces have been highly squeezed on these layers and have few features for detection. S^{3}FD [12] extends the range of layers to $conv3\_3$ with a 4 pixel stride to guarantee enough spatial features for tiny faces. Unfortunately, more anchor-associated layers give rise to another problem that the features of shallower layers are so semantically deficient that detectors fail to handle tiny faces in some complex situations.

To enrich semantics at all scales, FPN [14] and DSSD [15] adopt lateral connections to pass them from high-level deeper features to low-level shallower features in a top-down manner. However, we argue here that three problems may not be well addressed: **1) Semantic diffusion**. The top-down layer-by-layer transmission mode in FPN may cause semantics decay and even bring harmful noise[13], because the cues of tiny faces in high-level low-resolution features are rather weak (724 pixels receptive field in $conv7\_2$ [12]). This problem has been mentioned in [13], and we also conduct comparative experiments to reveal this problem. As shown in Fig. 1(a), the classification confidence maps of vanilla S^{3}FD, FPN and our method on $conv3\_3$ are visualized for tiny face detection. Compared with vanilla S^{3}FD and FPN, our method can not only cover more small faces (more bright points), but also obtain higher confidence scores (brighter points). For example, faces in the center region of the 1-st image, the faces near the top-right area of the 2-nd and the 3-rd image are undetected by FPN and S^{3}FD. **2) Static fusion strategy**. FPN combines different features in a simply linear combination without feature selection. [16] stated that the context is not always beneficial to face detection. For example, for a 10-pixel tall face, high-level features ($conv6\_2$, $conv7\_2$) with large receptive fields should not be overconcerned due to containing too much irrelevant information, as shown in Fig. 1.(b). The ideal fusing methods should be task-oriented where more relevant features should be boosted, while the irrelevant features should be suppressed. **3) Detecting face in one step**. Training a detector with the supervision of face area does not correspond with human visual perception because the information inside the tiny face is too little to capture a discriminative feature shown in Fig. 1.(b) It is intuitive that extending the region of the face to include more features is a feasible method as shown in Fig. 1.(c). Some works [1], [17] have also found that humans first locate a large and rough region and then refine the accurate position, especially in small objects.

To address the above three issues, this paper proposes an attention-guided semantically enriched feature aggregation architecture to create a feature pyramid with abundant semantics over all scales for face detection. Instead of a layer-by-layer transmission mode in FPN, a skip connection is used to directly pass high-level abstract information to shallower layers, which can relieve the problem of semantic diffusion. Furthermore, we propose a dynamic adaptive feature aggregation strategy based on attention mechanisms, which can be treated as a flexible and task-oriented feature selection where relevant



(a) Centralized Attention Feature (b) Distributed Attention Feature

Fig. 3. An illustration of our proposed CAF and DAF architectures. We present an example of 6-path feature aggregation with attention mechanisms. Notice that the upsampling layer is a simple bilinear resize operation in our experiments, which has the low computational complexity.

features are emphasized, and useless features are suppressed during feature fusion. To simulate the progressive learning style of humans, we propose a deep progressive refined loss (DPRL) that sequentially employs multi-scale supervisions in a coarse-to-fine manner. Based on the analysis above, we design different feature aggregation structures and conduct extensive experiments. Finally, two typical attention-guided feature aggregation structures are proposed: Centralized Attention Feature (CAF) and Distributed Attention Feature (DAF). Experiments have been carried on several public face detection benchmarks, which demonstrate that our proposed CAF and DAF are superior to the nave FPN for face detection. Fig. 2 shows an image containing a crowd of $1,000$ people with occlusion, scale, expression and illumination. Our methods successfully detect 845 faces.

The main contributions are summarized as follows,

- We investigate how to construct feature pyramids with enriched semantic and spatial information for face detection. Global attention is integrated into the aggregation framework, which learns to highlight relevant features while suppressing redundant ones.
- Inspired by the observation that the whole identification process of humans is from coarse to fine [1], we present a deep progressive refined loss to effectively facilitate feature learning. A large and rough region is located before finding an accurate face.
- Extensive experimental evaluations on four common face detection benchmarks, AFW, PASCAL face, FDDB and WIDER FACE datasets, demonstrate the superiority of the proposed framework.

II. Related Work

Face detection is a classical but challenging task that has received increasing attention in computer vision. Inspired by the successful applications of CNN in various problems [18], [8], state-of-the-art results for face detection have changed from handcrafted feature-based methods [19], [20], [21] to CNN-based methods [22], [23], [24], [25], [26]. Although CNN has promoted the growth of face detection, some challenging scenes containing large-scale variations of faces in practice have not been well addressed. We briefly categorize some recent works for handling multi-scale face detection into two classes: single-scale detector with image pyramid and multi-scale detector with feature pyramid.

A single-scale detector only utilizes a single layer in the CNN to detect objects on a specific scale. Faster RCNN [9] is a representative method that extracts scale-invariant features by region of interest (ROI) pooling. To remedy the coarse granularity of the last feature map in the backbone, [27], [28], [29] attempted to fuse features on different layers to obtain a semantically rich representation. Additionally, contextual information has been shown to be effective for face detection, especially for tiny or occluded faces. Body structure information has been successfully incorporated to facilitate estimating more accurate locations [29], [13]. In addition to a more powerful representation, more separate detectors are trained for each scale [1], and multi-level image pyramids are further utilized during inference. To reduce redundant image pyramids

and accelerate the detection pipeline, A scale aware network [30] is constructed to estimate face sizes in images and then build a customized and condensed image pyramid according to the estimated values. However, using image pyramids is very time-consuming because a set of images is required to pass a deep backbone many times during inference.

A multi-scale detector aims to learn multiple scale-specific representations in which different sizes of faces are independently detected by different feature maps. The seminal work of SSD [10] is based on this philosophy. SSD makes full use of inherently multi-scale features in the CNN and detects objects of various scales on distinct layers. Inherited from the SSD framework, carefully designed anchor strategies [12], [31] are introduced into detection layers that further improve the performance of finding tiny faces. ScaleFace [32] explicitly breaks down face detection into three subtasks according to face size: small, medium, and large. Three scale-variant detectors are designed to detect faces within a certain range of scales. Similarly, SSH [11] is also capable of detecting various faces in a single forward pass, where three detection modules are jointly trained from the convolutional layers with different strides. Because those methods only use an individual layer in a network for prediction, it is difficult for low-level features with weak semantics to handle tiny and occluded faces. Recent works [14], [15], [33], [34] show that combining fine-grained details and high-level semantics is beneficial for providing a robust prediction. FPN [14] and DSSD [15] adopt a lateral connection to fuse adjacent layers in a top-down manner. Experiments demonstrate that semantically augmented features can improve the performance of all ranges of faces. Based on FPN, RON [33] uses more sophisticated connections to learn strong representations.

III. The Proposed Approach

In this section, we present our attention-guided semantically enriched feature aggregation framework to learn a semantic feature pyramid for face detection. The overall architecture is introduced first, and then, two typical feature aggregation structures and DPRL are described in detail.

A. Overall Architecture

Face detection requires both spatial details and abstract information for face or nonface probability estimation and bounding box regression. SSD-based methods depend on high-resolution but semantically weak features to detect tiny faces, making the face detection pipeline suboptimal. A feasible way is to transmit high-level semantics to all detection layers. Inspired by ResNet [35], U-Net [36] and FPN [14], we employ skip connections to directly pass multi-level semantic information to different layers, aiming at preserving as much as possible during transmission.

In addition, recent work shows that introducing gates into the standard model can facilitate optimization and improve the performance [37], [38]. The attention mechanism is generally employed as a gate, which has been widely used in machine translation [39] and visual recognition [40], [41]. Therefore, it is intuitive to embed attention into our model to control the aggregation of different layers. A reasonable explanation is that irrelevant regions (or channels) are implicitly suppressed while more relevant features are highlighted during feature aggregation. When detecting extremely small faces, high-level features in the top layer may be very weak and contain a mostly irrelevant background. For example, the receptive field in $conv7\_2$ is 724 pixels that is much greater than 10-pixel tall faces.

In summary, the proposed framework consists of three key components: multi-level semantic feature aggregation, attention-based gates and DPRL. We introduce two typical feature aggregation structures as shown in Fig. 3: a Centralized Attention Feature (CAF) and a Distributed Attention Feature (DAF).

B. Centralized Attention Feature

The CAF first learns a feature tower that combines a multi-level context by spanning hierarchical features with element-wise sum aggregation. Benefiting from these operations, the feature tower has discriminability and contains abundant multi-scale semantic information, which leads to stronger feature pyramids for various scales of faces, especially tiny faces. Fig. 3 shows an example of the proposed CAF with 6-path feature aggregation.

Feature hierarchy in the backbone. Given a single image x, we denote the CNN feature extraction process as $[\phi_1, \phi_2, ..., \phi_L] = Conv(x)$, where $Conv(.)$ is defined by the convolution neural network with L prediction layers for detection in the backbone, and ϕ_i is the output of i-th prediction layer. Assume that VGG16 is a backbone where $[\phi_1, ..., \phi_L]$ corresponds to $[conv3\_3, conv4\_3, conv5\_3, fc7, conv6\_2, conv7\_2]$. SSD adopts multiple feature maps as the prediction layers / anchor-associated layers Φ_{pred}.

$$\Phi_{pred} = \{\phi_p, \phi_{p+1}, ..., \phi_q\}, \tag{1}$$

where $p = q = L$ in Faster RCNN, $p = 2$ ($conv4\_3$) and $q = L$ in SSD and $p = 1$ ($conv3\_3$) and $q = L$ in S^{3}FD. Generally, the shallower layers with higher resolution are used to detect the small objects, and deeper layers are used to detect large objects.

Feature aggregation with attention. To enrich semantics in shallower layers and augment details in deeper layers simultaneously, CAF creates a feature tower that aggregates multi-level representations.

$$\Psi_{tower} = Agg(\phi_p, \phi_{p+1}, ..., \phi_q), \tag{2}$$

where $Agg(.)$ is the aggregation function of the prediction layers. In the implementation of $Agg(.)$, we simply upsample the low-resolution layers and then concatenate them. $q-p+1$ is the number of layers for aggregation. As shown in [1], the contributions of diverse level feature maps for detecting a certain range of faces are different. To promote more relevant features and suppress useless features, an attention mechanism is imposed into $Agg(.)$. Thus, feature aggregation can be further formulated as follows:

$$\Psi_{tower} = Agg(\phi_p, ..., \phi_q, Att(\phi_p, ..., \phi_q)), \tag{3}$$

where $Att(.)$ is the attention function. In this paper, we present a global cross-layer channel-wise attention that is expansion of vanilla channel-wise attention [40]. A mathematical description of the attention function is presented as follows:

$$Att(\phi_p, ..., \phi_q) = [W_p, ..., W_q], \tag{4}$$

where $W_p \in R^{c_p}$ has the same number of channels c_p as ϕ_p. $Att(.)$ generates a vector $[W_p, ..., W_q] \in R^{c_p+...+c_q}$ as weights for each channel of all layers. We adopt the SEBlock in [40], which consists of *squeeze* and *excitation*. The first stage *squeeze* imposes a global pooling operation on each channel of ϕ_p. For k^{th}-channel $\phi_p^k \in R^{W \times H}$ in ϕ_p:

$$z_p^k = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} \phi_p^k(i, j) \tag{5}$$

where $\phi_p^k(i, j)$ is the element on the k^{th} channel, i^{th} column and j^{th} row. The *excitation* stage contains two fully-connection layers fc_2 and a sigmoid function to obtain the attention weights on each channel with input z_p:

$$W_p = sigmoid(fc_2(z_p)) \tag{6}$$

Finally, feature aggregation $Agg(.)$ rescales the input features of all layers:

$$\Psi_{tower} = Agg(\phi_p, ..., \phi_q, Att(\phi_p, ..., \phi_q)), \tag{7}$$
$$= [W_p \otimes \phi_p, ..., W_q \otimes \phi_q] \tag{8}$$

where $\otimes$ denotes channel-wise multiplication. A more detailed desciption can be found in [40].

The weights across different channels among all layers are learned in a global view, which makes the feature aggregation dynamic and self-adaptive. Similar to feature selection, relevant features are emphasized and the useless features are suppressed.

The attention function $Att(.)$ in our model is flexible. It can be channel-wise [40] or pixel-wise [41] attention. Here, we take the proposed global cross-layer channel-wise attention as an example. The motivation is to learn to select more relevant features instead of learning the dependent relation between layers.

Feature pyramid generation. Our target is to generate a feature pyramid for classification and bounding box regression. After feature fusing with attention, the obtained feature tower contains multi-scale features with multi-level semantics, leading to high discriminative abilities, especially for tiny faces. CAF generates a new feature hierarchy based on the feature tower representation Ψ_{tower}.

$$\phi_i' = T_i(\Psi_{tower}) \qquad i = p+1, p+2, ..., q, \tag{9}$$

where ϕ_i' denotes the generated i-th level representation. $T_i(.)$ is the transformation function for the i-th level. In this paper, multiple convolution operations with stride=2 are utilized to generate the feature pyramid.

C. Distributed Attention Feature

Different from CAF, DAF generates multiple towers with different resolutions composed of feature pyramids by aggregating multi-level features.

For the i-th layer of feature hierarchy in backbone, different hierarchies of features are generated through a series of ampling functions. Notice that the upsampling layer is a simple bilinear resize operation in our experiments, which has low computational complexity.

$$\phi_i^k = \mathcal{U}_{\times 2}(\phi_i^{k-1}) \qquad k = 2, 3, .., q, \tag{10}$$
$$\phi_i^k = \phi_i \qquad k = 1, \tag{11}$$

where $\mathcal{U}_{\times 2}(.)$ is the upsampling function to extend the feature map by 2 times, and k is the hierarchy of the feature maps.

To learn a feature tower for the j-th level in the feature pyramid, all features after the j-th layer are combined with the j-th features. Attention is also applied such as CAF.

$$\Psi_{tower}^j = Agg(\phi_j, ..., \phi_q^{q-j+1}, Att(\phi_j, ..., \phi_q^{q-j+1})), \tag{12}$$

where $\Psi_{tower}^j (j \in \{p, ..., q\})$ is the aggregated feature tower that is used as prediction layer. Fig. 3 depicts the DAF with 6 paths for aggregation, i.e., $p = 1, q = 6$.

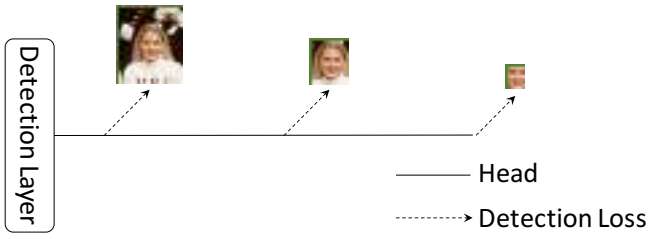


Fig. 4. The deep progressive refined loss (DPRL) sequentially employs different scopes of supervision from coarse to fine.

D. Deep Progressive Refined Loss

At present, most face detectors utilize face regression and classification as the only loss function, where there is no supervision in the middle layers. The singleness of supervision affects the detection of faces with different scales. This issue seems unimportant in normal and large face detection because they have adequate pixels and salient features. Nevertheless, little information is available to represent small faces as shown in Fig. 1(b) and (c). Their features are further squeezed with the increase of stride. Consequently, finding tiny faces based on few features is difficult.

To address this issue, we propose a deep progressive refined loss (DPRL) to utilize more context beyond the face extent. Inspired by deeply supervised learning [42], we sequentially employ different scopes of supervision from coarse to fine. The structure of DPRL is shown in Fig. 4. Our DPRL adopts a three-branch structure, where a shallower branch is used to locate a larger region (including the head and shoulder) and a deeper branch is used to locate a finer face. This learning process from easy to difficult is similar to curriculum learning [43], which is consistent with human learning style in a meaningful order. In particular, we design three sets of

anchors with different sizes that correspond to three scopes of face (large, middle and normal region). The Multi-task losses [9], [12] including classification loss (softmax) and bounding box regressions are defined as follows:

$$L(\{p_i\},\{t_i\}) = \frac{\lambda}{N_{cls}}\sum_i L_{cls}(p_i,p_i^*) + \frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i,t_i^*),$$

where i is the index of the anchor and p_i is the predicted classification probability of anchor i. p_i^* is the ground-truth label of anchor i, where $i = 0$ represents a negative anchor and $i = 1$ represents a positive anchor. t_i is a vector representing the 4 parameterized coordinates of the predicted bounding box and t_i^* is the ground truth coordinates. Softmax loss is adopted in $L_{cls}(p_i, p_i^*)$. The bounding box regression loss $L_{reg}(t_i, t_i^*)$ is the smooth L_1 loss defined in [44]. N_{cls} and N_{reg} are normalized terms, and λ is a balancing parameter.

DPRL contains three sets of anchors with different sizes, i.e., large anchors la_i, middle anchors ma_i and normal anchors na_i. Their predicted classification probabilities of anchor i are lp_i, mp_i, np_i, and the coordinates of the predicted bounding box are lt_i, mt_i, nt_i. Thus, DPSL is formulated as follows:

$$\begin{aligned} L_{DPRL} = {} & L_l(\{lp_i\},\{lt_i\}) + L_m(\{mp_i\},\{mt_i\}) \\ & + L_n(\{np_i\},\{nt_i\}) \end{aligned} \tag{13}$$

Notice that we only use the output of normal anchors in inference, which means no additional computational cost is introduced.

E. Training

In the following sections, we describe the implementation details including the training dataset, data augmentation and other significant techniques for the overall framework training.

1) Training data and data augmentation.: All models are trained on the WIDER FACE training dataset [45]. Following data augmentation in [10], [12], color distortion and horizontal flipping are adopted in this paper. A random crop is also used to generate training images [12]. In addition, we utilize data-anchor-sampling [13] in our final face detector to compare with various state-of-the-art detectors.

2) Hard negative mining.: To alleviate a significant imbalance between positive and negative samples after anchor matching, an online hard negative mining strategy [10] is employed during training. Specifically, all negative anchors are sorted by the classification loss, and the top ones are selected for training to ensure that the ratio between negative and positive anchors is at most $3:1$.

3) Other implementation details.: We use VGG16 and ResNet-50 as backbone networks, which are initiated by the pre-trained model [46][35]. The newly added layers are initialized with Xavier. To obtain the same resolution of layers for aggregation, a simple bilinear resize operation is utilized as an upsampling method. The sizes of large and middle anchors are $2.0\times$ and $1.5\times$ than those of the normal anchors. We use mini-batch SGD with the momentum of 0.9 and weight decay of 0.0005. The batch size is set to 12 on four GPUs. The initial learning rate is 0.001 and decreases 10 times at iteration $80k$ and $100k$, and the training ends at $120k$ iterations.

mAP(%) Methods \ Subsets	Easy	Medium	Hard
S^{3}FD + CF / CAF-3path	94.2 / 94.2	92.9 / 92.9	86.5 / 86.6
S^{3}FD + CF / CAF-4path	94.4 / 94.4	92.9 / 93.0	86.6 / 86.6
S^{3}FD + CF / CAF-5path	94.2 / 94.3	92.8 / 93.1	86.4 / 86.7
S^{3}FD + CF / CAF-6path	94.0 / 94.3	92.9 / 93.0	86.4 / 86.6
S^{3}FD + DF / DAF-3path	94.4 / 94.4	93.2 / 93.3	86.7 / 86.7
S^{3}FD + DF / DAF-4path	94.4 / 94.4	93.1 / 93.2	86.4 / 86.9
S^{3}FD + DF / DAF-5path	94.2 / 94.4	93.0 / 93.1	86.4 / 86.6
S^{3}FD + DF / DAF-6path	94.0 / 94.2	92.9 / 93.1	86.4 / 86.5

TABLE I
THE COMPARATIVE RESULTS OF OUR CF AND DF ALONG WITH THEIR ATTENTION VERSIONS, CAF AND DAF ON WIDER FACE VALIDATION SUBSET. XXX-nPATH MEANS THAT THE FIRST n PREDICTION LAYERS ARE USED FOR FEATURE AGGREGATION.

IV. EXPERIMENT

In this section, an ablation study is first conducted to analyze the effectiveness and significance of each part in our methods, including different feature aggregation structures, attention-based gates and DPRL, and then we evaluate the final models on four widely used face detection benchmarks.

A. Model analysis

We conduct extensive experiments on the WIDER FACE validation set to analyze our model. The WIDER FACE validation dataset is a comprehensive and challenging dataset that contains easy, medium and hard levels.

1) Baseline: Our proposed CAF and DAF are general-purpose and applicable to arbitrary detection models, so we simply adopt S^{3}FD [12] as a baseline in this paper. Because our method is inspired by FPN [14], S^{3}FD and FPN are combined as a baseline (S^{3}FD + FPN) to compare with our methods.

Inference time. We measure the speed of FPN and our DAF with batch size 1 on a Tesla P40, CUDA 8.0 and cuDNN v6. For the 640x384 input size, FPN runs at 21 PFS, while our DAF achieves 28 FPS. Our DAF is 7 FPS faster than FPN. The underlying reasons include two aspects: 1) DAF uses fewer channels (256 for all layers) than FPN (same as the backbone) during feature aggregation. 2) The strategies employed in the training stage do not increase the computational complexity in reference, such as DPRL. Thus, our DAF achieves both faster running speed and higher performance.

2) Ablation study: We perform an ablation study to validate the efficacy of each proposed component in this paper, as well as various choices in designing the network. Our model is evaluated under three different settings: (i) CF and DF that mean CAF and DAF w/o attention: they only employ our proposed frameworks to aggregate feature maps without attention mechanisms. (ii) CF n-path and DF n-path: the first n layers are used to aggregate multi-level feature maps. (iii) CAF n-path and DAF n-path: attention is incorporated when aggregating n layers.

Tab. I and Tab. II show the results of different methods under different settings. Some promising conclusions can be summarized as follows:

TABLE II
THE COMPARATIVE RESULTS OF OUR METHODS WITH DIFFERENT BACKBONE NETWORKS AND BASELINES (VANILLA S^{3}FD AND FPN) ON THE WIDER FACE VALIDATION SUBSET.

Backbone	Method	AP		
		Easy	Medium	Hard
VGG-16	S^{3}FD	94.0	92.7	84.2
	S^{3}FD + FPN	93.9	92.9	85.9
	S^{3}FD + CF	94.4	92.9	86.6
	S^{3}FD + CAF	94.3	93.1	86.7
	S^{3}FD + DF	94.4	93.2	86.7
	S^{3}FD + DAF	94.4	93.2	86.9
ResNet-50	S^{3}FD	94.5	93.1	85.1
	S^{3}FD + FPN	94.5	93.6	87.0
	S^{3}FD + CF	95.1	93.9	88.4
	S^{3}FD + CAF	95.5	94.1	88.7
	S^{3}FD + DF	95.3	93.9	88.5
	S^{3}FD + DAF	95.6	94.3	88.9

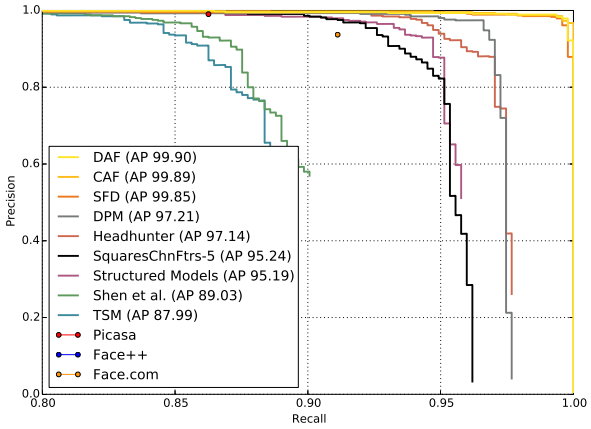


Fig. 5. Evaluation on the AFW dataset.

Low-level feature aggregation for CF and DF is crucial for detecting hard faces. We conduct some experiments to explore whether the more layers of fusion, the better the effect we obtain. From Tab. I, we can see that the performance of CF is improved with the number of layers increasing in the beginning. However, when aggregating more layers, the performance no longer increases or even declines. Our DF shows similar behaviors. Finally, CF-4path and DF-3path achieve the best results. This phenomenon may be attributed to two reasons: 1) features with a large gap in scale may not help each other, which also exists in FPN [13]; 2) other scales of features may dominate the main scale. For example, detecting tiny faces mainly depends on $conv3\_3$, but incorporating more scales of features may drown out the effect of $conv3\_3$. Therefore, introducing gates to adaptively modulate the aggregation of different layers is significant.

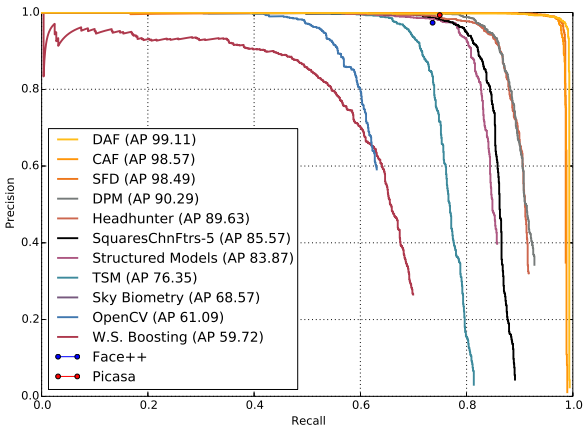
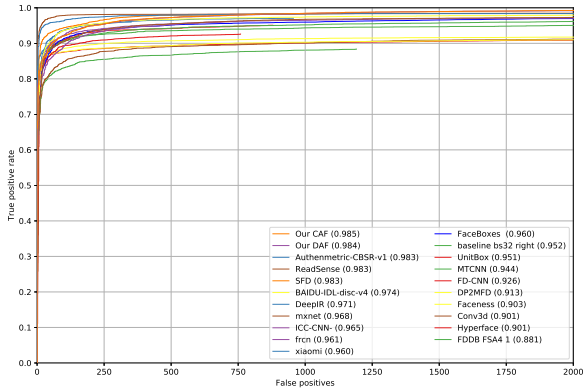
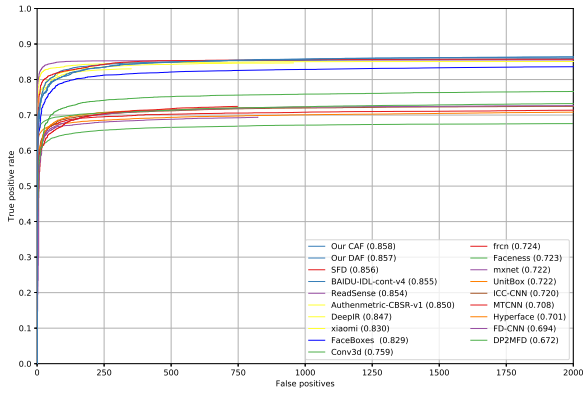


Fig. 6. Evaluation on the PASCAL face dataset.



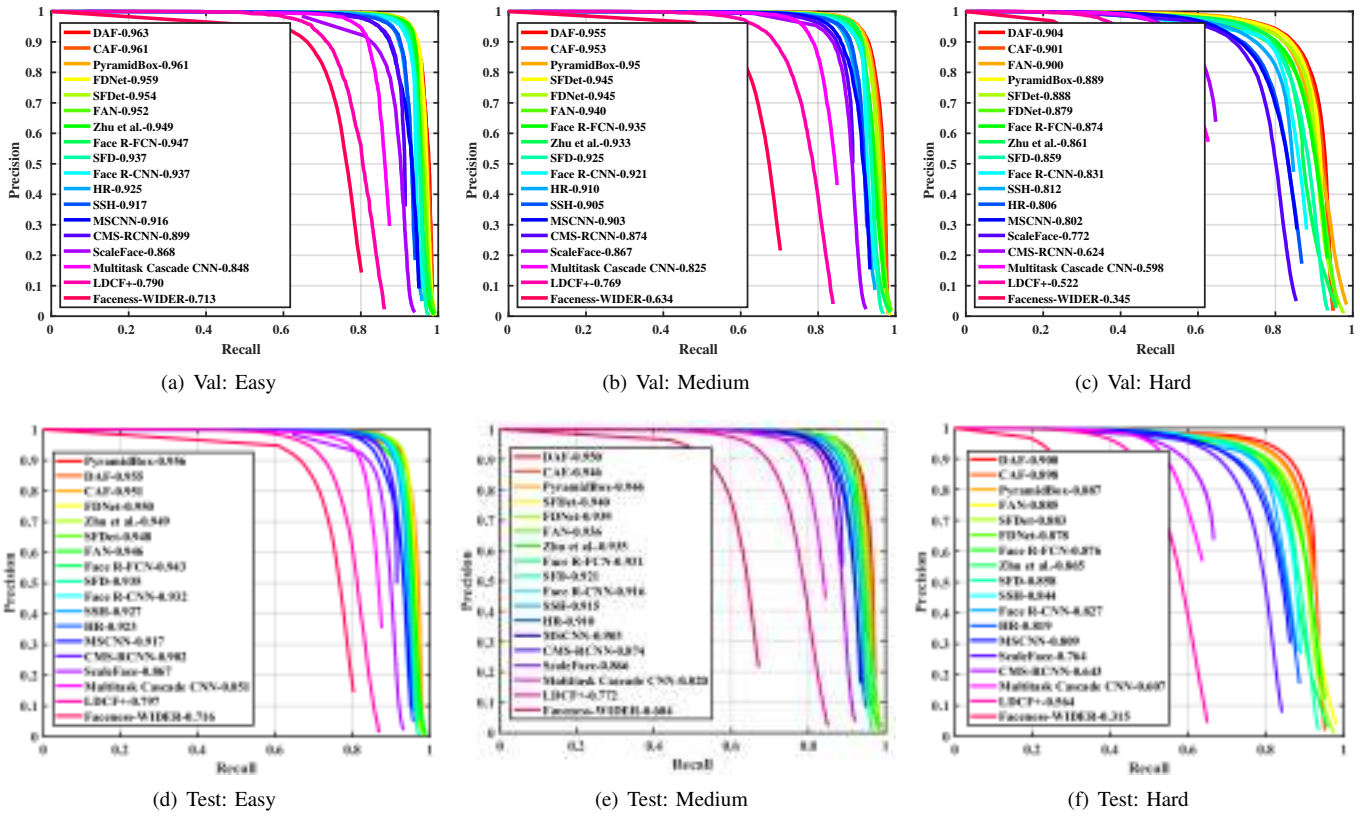
(a) Discontinuous ROC curves



(b) Continuous ROC curves

Fig. 7. Evaluation on the FDDB dataset.

Attention-based gates boost the performance of detection. We add an attention mechanism to our CF and DF. Tab. I shows that the performances of all the methods are improved when attention is introduced. It is interesting to observe that there is only a small improvement in the small number of aggregation layers (such as 3-path and 4-path). With increasing aggregation layers, the channel-wise attention mechanism highlights its importance. Most models have approximately $0.2 \sim 0.3\%$ improvement (such as CAF/DAF-6path on the easy subset, CAF/DAF-5path on the hard subset), which means



(a) Val: Easy (b) Val: Medium (c) Val: Hard

(d) Test: Easy (e) Test: Medium (f) Test: Hard

Fig. 8. Precision-recall curves on the WIDER FACE validation set and testing set.

that attention is more beneficial for large layer aggregation and is mainly because large amount of layers require dynamic modulation by attention. These results further validate the effectiveness of attention.

The structures of CAF and DAF are beneficial for detection. Tab. II shows the mAP of the baselines for different backbone networks and our proposed methods. Methods with ResNet-50 as the backbone consistently perform better than those with VGG-16 as the backbone. Due to the deficiency semantics on shallow layers, S^{3}FD yields inferior performance at all scales of faces compared to FPN, especially for the hard setting from 84.2% to 85.9% on VGG-16 and from 85.1% to 87.0% on ResNet-50. Our proposed two structures (CF, DF, CAF and DAF) can be treated as an alternative method to FPN. The experimental results clearly show that all of our proposed methods significantly outperform S^{3}FD, especially on small faces (increasing by 2.7% and 4.8% on VGG-16 and ResNet-50 respectively). To demonstrate the weak semantic problem for tiny face detection, we visualize the classification confidence maps on $conv3\_3$ of S^{3}FD and our method. From Fig. 1, we can see that our approach can not only cover more tiny faces than S^{3}FD but also obtain higher confidence. When incorporating the attention mechanism into our models, the performance of face detection can be further improved; especially 1.0% and 1.9% improvement are achieved compared to FPN in the hard set. This underlines that CAF and DAF can further enhance the semantics on low-level feature maps (like $conv3\_3$, $conv4\_3$), which make it more robust to tiny faces. In addition, the performance on the medium and easy set is also better than FPN, which validates the effectiveness of CAF and DAF.

DPRL is beneficial for face detection. We investigate various configurations of DPRL, including two granularities, large and middle scopes. DAF is our baseline, mc and lc represent middle and large scopes of supervision. p means deploying supervisions in parallel. As shown in Table III, employing a middle level mc boosts the performance. When two scopes are introduced sequentially, the performance is further improved with 0.7%, 0.7% and 1.1% AP improvements on VGG-16. When the backbone changes to ResNet-50, the improvement is more obvious. Additionally, experiments show that the sequential learning strategy performs better than the parallel strategy. The underlying reasons are that this coarse-to-fine training strategy is helpful for learning discriminative features. In addition, locating rough regions is easier and can narrow the search range for tiny face detection. In addition, we also evaluate the impact of DPRL on S^{3}FD + FPN, and Table III shows that DPRL boosts its performance by 1.1%, 1.0% and 1.4% on VGG-16 and 0.8%, 1.0% and 1.9% on ResNet-50.

B. Evaluation on Benchmarks

We evaluate our proposed CAF and DAF on four face detection benchmarks, including Annotated Faces in the Wild (AFW)[47], PASCAL Face[48], the Face Detection Dataset

and Benchmark (FDDB) [49] and WIDER FACE [45]. For a fair comparison, we follow the testing protocols in S3FD[12]. All methods are only trained on the WIDER FACE training dataset and are directly tested on these four face detection benchmarks without fine-tuning.

1) AFW Dataset: It has 205 images with 473 labeled faces, which were collected from Flickr. Its challenge is due to large variations in appearance and viewpoint. To verify the performance of the proposed methods, we evaluate them against other works, including deep learning methods, traditional methods and some commercial face detectors.

The commercial face detectors include Face.com, Face++ and Picasa. For traditional methods, we compare the Deformable Parts Model (DPM)[50], Headhunter [20], SquaresChnFtrs-5, Structured Models [48], Shen et al. [51], and TSM [47]. The deep learning methods contain S3FD [12]. All of the compared methods are from their released results.

Fig. 5 shows the precision-recall curves of different methods. We make the following observations. Due to the powerful representation of CNN, three deep learning methods (S3FD, CAF and DAF) are significantly better than the traditional face detectors. The improvements are particularly apparent when the recall is high. Although S3FD achieves a near-perfect performance, our CAF and DAF further improve AP from 99.85% to 99.89% and from 99.85% to 99.90%. Some qualitative results on the AFW dataset are presented in Fig.9. Note that the state-of-the-art methods on this dataset have performed well, and even a small improvement is difficult. These results highlight the effectiveness of our methods.

2) PASCAL face Dataset: It contains $1,335$ labeled faces in 851 images with large face appearance and pose variations. They were collected from the test set of the PASCAL person layout dataset. Thus, it is a subset of PASCAL VOC[52]. Compared to the AFW dataset, the PASCAL face contains more images under various conditions. We compare our methods with commercial face detectors (Picasa, Face++ and SkyBiometry) and other face detectors (DPM[50], Headhunter [20], SquaresChnFtrs-5, Structured Models [48] and TSM [47]).

The precision-recall curves of different methods are shown in Fig. 6. The three deep learning methods (S^3FD, CAF and DAF) perform significantly better than the traditional methods. Compared with DPM, S^3FD raises performance from 90.29% to 98.49%. Although S^3FD has achieved very high performance, our methods further improve the AP from 98.49% to 99.11%, which is mainly because our CAF and DAF can build a semantic-enriched feature pyramid, which is effective in dealing with hard face detection. We also present some qualitative results on the PASCAL face dataset in Fig.10.

3) FDDB Dataset: The Face Detection Data Set and Benchmark (FDDB) [49] is a well-known benchmark that contains $5,171$ faces in $2,845$ images collected from the Yahoo! News website. It is a challenging dataset due to arbitrary poses, occlusions, various lighting, expressions, low-resolution and out-of-focus faces. All faces in this dataset are annotated with ellipses. Following the evaluations in [49], there are two metrics based on ROC. The discrete score metric is similar to previous evaluations, which is a coarse match between prediction and the ground truth. Another metric is a precise one.

TABLE III
EFFECTIVENESS OF DPRL ON THE AP PERFORMANCE.

Backbone	Method	AP		
		Easy	Medium	Hard
VGG-16	S^3FD + FPN	93.9	92.9	85.9
	S^3FD + FPN-mc-lc	94.5	93.4	87.1
	DAF	94.4	93.2	86.9
	DAF-mc	95.1	93.8	87.6
	DAF-mc-p	94.8	93.6	87.3
	DAF-mc-lc	95.1	93.9	88.0
ResNet-50	S^3FD + FPN	94.5	93.6	87.0
	S^3FD + FPN-mc-lc	95.3	94.6	88.9
	DAF	95.6	94.3	88.9
	DAF-mc	95.9	94.9	89.7
	DAF-mc-p	95.8	94.7	89.4
	DAF-mc-lc	96.3	95.5	90.4

We compare the proposed CAF and DAF with other published state-of-the-art methods, including [17], [1], [53], [25], [54], [55], [56]. The results of the compared methods are from the FDDB website[1]. The ROC curves of the discrete score metric and the continuous score metric are depicted in Fig. 7(a) and Fig. 7(b), respectively. It can be observed that our CAF and DAF achieve state-of-the-art performance and outperform others on discontinuous and continuous ROC curves. Note that the FDDB dataset uses ellipses as the ground truth of face. Although several methods also predict special detections, our methods still perform better. Fig.11 shows several qualitative results on the FDDB dataset. The results demonstrate that the proposed methods are robust to various scales, extreme pose (profile face), heavy occlusion and blur conditions.

4) WIDER FACE Dataset: The WIDER FACE Dataset [45] is widely used in face detection evaluations because it is the most challenging face detection database. It contains 393,703 faces in 32,203 images with a high degree of variability in scale, pose, illumination and occlusion. It has three levels of difficulty (easy, medium and hard) according to the difficulty of detection. Samples in the dataset are split into training (40%), validation (10%) and testing (50%) sets. Our CAF and DAF are trained only on the training set and tested on the validation set and the testing set. We compare the proposed methods with recent state-of-the-art methods, including Zhu. et al.[31], MSCNN[57], Face R-FCN[55], S^3FD[12], SSH[11], ScaleFace[32], Face R-CNN[58], HR[1], LDCF+[54], CMS-RCNN[29], Multitask Cascade CNN[25], Faceness-WIDER[24], ACF-WIDER[53], Two-stage CNN[45], Multiscale Cascade CNN[45], FAN[58], FANet[59], SFDet[60], FDNet[61] and Pyramidbox[13]. The Precision-Recall (PR) curves and average precisions (AP) are

[1]http://vis-www.cs.umass.edu/fddb/results.html

Fig. 9. Qualitative results on the AFW dataset. The green bounding box represents the detector confidence above 0.8.

Fig. 10. Qualitative results on the PASCAL face dataset. The green bounding box represents the detector confidence above 0.8.

used for comparison among different methods. Fig. 8 shows the results on both validation and testing datasets. We make the following observations:

The performances of all methods follow the difficulties of three subsets where they perform the best on the easy subset. The first row of Fig. 8 is the results on the validation set. Although the baseline model S^3FD obtains 93.7% AP on the easy subset, our CAF and DAF still outperform it by approximately 2.4% and 2.6%. In addition, our methods yield superior performance on the medium subset and the performance gap have been further increased (3.0% AP increase compared with S^3FD), which demonstrates that our methods are also beneficial for detecting middle-scale faces. The main reason is that the low-level layers with enough spatial details can be spread directly to the high-level layers. Hence, high-level features have more abundant information to accurately detect bigger faces. On the hard subset, our CAF and DAF also achieve state-of-the-art performance. In particular, DAF and CAF outperform S^3FD by approximately 4.2% and 4.5% on the validation set, respectively. The second row shows the results on the testing set. Since the testing set contains more samples with a large variance of scale, pose, occlusion, etc, all methods perform inferior on it. On the easy subset, our proposed CAF and DAF almost surpass all other methods but are marginally inferior to Pyramidbox (only 0.1% gap). The underlying reason is that our methods focus on improving the performance on hard face detection, especially for tiny faces. Thus, our proposed methods achieve state-of-the-art performances on the medium and hard subsets of the testing sets. Particularly, our DAF and CAF outperform the S3FD by 2.0%/1.6%, 2.9%/2.5% and 4.5%/4.3% on three subsets of the testing set. It further validates the generalization ability of our models.

Fig.12 and Fig. 13 show some examples of detected faces in the WIDER FACE dataset by our methods. Many tiny faces in very crowded scenes have been successfully detected. Our face detector can also address other hard conditions, such as variations in pose, occlusion, expression, makeup and illumination.

V. Conclusion

In this paper, we propose a novel feature aggregation framework based on attention gates for face detection. Two typical structures named CAF and DAF are constructed to learn a feature pyramid with semantics at all layers, which are more effective for tiny faces. In addition, attention has

Fig. 11. Qualitative results on the FDDB dataset. The green bounding box represents the detector confidence above 0.8.

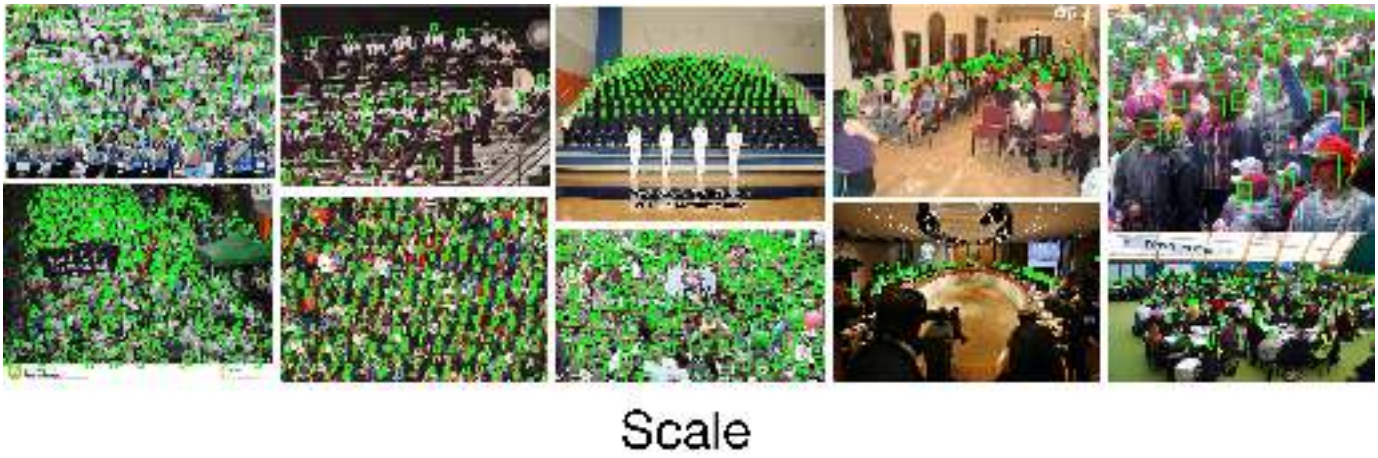


Fig. 12. Our methods can handle faces with a wide range of face scales. The green bounding box represents the detector confidence above 0.8.

been studied to adaptively control the information flow of each layer during feature aggregation. DPRL, which utilizes more context, is presented to detect faces in a coarse-to-fine manner. Experimental results across four structures on challenging face detection databases show that our CAF and DAF significantly outperform state-of-the-art face detection methods. In our future work, we intend to extend our methods to generic object detection.

VI. Acknowledgment

The authors would like to sincerely thank the associate editor and the reviewers for their valuable comments and advice. This work was supported in part by the State Key Development Program under Grant 2016YFB1001001, in part the National Natural Science Foundation of China under Grant 61622310, and in part by the Beijing Natural Science Foundation under Grant JQ18017.

References

[1] P. Hu and D. Ramanan, "Finding tiny faces," in *CVPR*, 2017.
[2] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *CVPR*, 2008.
[3] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *CVPR*, 2013.
[4] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *CVPR*, 2016.
[5] O. M. Parkhi, A. Vedaldi, A. Zisserman *et al.*, "Deep face recognition." in *BMVC*, 2015.
[6] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *TIFS*, 2018.
[7] 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," *IJCV*, 2015.

Fig. 13. Our methods are robost to makeup, illumination, pose, occlusion, expression and blur. The green bounding box represents the detector confidence above 0.8.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.

[9] S. Ren, K. He, R. G, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. B, "Ssd: Single shot multibox detector," in *ECCV*, 2016.

[11] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis, "Ssh: Single stage headless face detector." in *ICCV*, 2017.

[12] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "Sˆ 3fd: Single shot scale-invariant face detector," in *ICCV*, 2017.

[13] X. Tang, D. K. Du, Z. He, and J. Liu, "Pyramidbox: A context-assisted single shot face detector," in *ECCV*, 2018.

[14] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection." in *CVPR*, 2017.

[15] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "Dssd: Deconvolutional single shot detector," *arXiv preprint arXiv:1701.06659*, 2017.

[16] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware cnn model," in *ICCV*, 2015.

[17] K. Zhang, Z. Zhang, H. W, Z. Li, Y. Qiao, and W. Liu, "Detecting faces using inside cascaded contextual cnn," in *ICCV*, 2017.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[19] P. Viola and M. J. J, "Robust real-time face detection," *IJCV*, 2004.

[20] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *ECCV*, 2014.

[21] S. Liao, A. K. Jain, and S. Z. Li, "A fast and accurate unconstrained face detector," *TPAMI*, 2016.

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

[23] H. Qin, J. Yan, X. Li, and X. Hu, "Joint training of cascaded cnn for face detection," in *CVPR*, 2016.

[24] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *ICCV*, 2015.

[25] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *SPL*, 2016.

[26] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "Unitbox: An advanced object detection network," in *ACMMM*, 2016.

[27] S. Bell, C. Lawrence Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *CVPR*, 2016.

[28] T. Kong, A. Y, Y. C, and F. S, "Hypernet: Towards accurate region proposal generation and joint object detection," in *CVPR*, 2016.

[29] C. Zhu, Y. Zheng, K. Luu, and M. Savvides, "Cms-rcnn: contextual multi-scale region-based cnn for unconstrained face detection," in *Deep Learning for Biometrics*, 2017.

[30] Z. Hao, Y. Liu, H. Qin, J. Yan, X. Li, and X. Hu, "Scale-aware face detection," in *CVPR*, 2017.

[31] C. Zhu, R. Tao, K. Luu, and M. Savvides, "Seeing small faces from robust anchors perspective," in *CVPR*, 2018.

[32] S. Yang, Y. Xiong, C. C. Loy, and X. Tang, "Face detection through scale-friendly deep convolutional networks," *arXiv preprint arXiv:1706.02863*, 2017.

[33] T. Kong, F. Sun, A. Yao, H. Liu, M. Lu, and Y. Chen, "Ron: Reverse connection with objectness prior networks for object detection," in *CVPR*, 2017.

[34] T. Kong, F. Sun, C. Tan, H. Liu, and W. Huang, "Deep feature pyramid reconfiguration for object detection," in *ECCV*, 2018.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[36] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[37] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, 1997.

[38] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[40] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.

[41] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *CVPR*, 2017.

[42] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artificial Intelligence and Statistics*, 2015.

[43] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *ICML*, 2009.

[44] R. Girshick, "Fast r-cnn," in *ICCV*, 2015.

[45] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *CVPR*, 2016.

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[47] D. Ramanan and X. Zhu, "Face detection, pose estimation, and landmark localization in the wild," in *CVPR*, 2012.

[48] J. Yan, X. Zhang, Z. Lei, and S. Z. Li, "Face detection by structural models," *Image and Vision Computing*, 2014.

[49] V. Jain and E. Learned-Miller, "Fddb: A benchmark for face detection in unconstrained settings," Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, Tech. Rep., 2010.

[50] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *TPAMI*, 2009.

[51] X. Shen, Z. Lin, J. Brandt, and Y. Wu, "Detecting and aligning faces by image retrieval," in *CVPR*, 2013.

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

[53] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Aggregate channel features for multi-view face detection," in *IJCB*, 2014.

[54] E. Ohn-Bar and M. M. Trivedi, "To boost or not to boost? on the limits of boosted trees for object detection," in *ICPR*, 2016.

[55] Y. Wang, X. Ji, Z. Zhou, H. Wang, and Z. Li, "Detecting faces using region-based fully convolutional networks," *arXiv preprint arXiv:1709.05256*, 2017.

[56] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "Faceboxes: A cpu real-time face detector with high accuracy," in *IJCB*, 2017.
[57] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *ECCV*, 2016.
[58] H. Wang, Z. Li, X. Ji, and Y. Wang, "Face r-cnn," *arXiv preprint arXiv:1706.01061*, 2017.
[59] J. Zhang, X. Wu, J. Zhu, and S. C. Hoi, "Feature agglomeration networks for single stage face detection," *arXiv preprint arXiv:1712.00721*, 2017.
[60] S. Zhang, L. Wen, H. Shi, Z. Lei, S. Lyu, and S. Z. Li, "Single-shot scale-aware network for real-time face detection," *IJCV*, 2019.
[61] C. Zhang, X. Xu, and D. Tu, "Face detection using improved faster rcnn," *arXiv preprint arXiv:1802.02142*, 2018.

Zhihang Li received his B.S. degree from the School of Information Engineering, China University of Geosciences in 2016. He is currently pursuing his Ph.D. degree with the National Laboratory of Pattern Recognition (NLPR), Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. His research interests include deep learning, computer vision, biometrics, and machine learning.

Xu Tang received his B.S. degree from Hefei University of Technology and his M.S. degree from the University of Chinese Academy of Sciences in 2014 and 2017, respectively. He has been working on computer vision at Baidu, Inc. since 2017. He is now a Senior R&D Engineer of the computer vision team at Baidu, Inc. His research interests focus on face detection and recognition, tracking.

Xiang Wu received his B.E. and M.S. degrees in electronic engineering from the University of Science and Technology Beijing in 2013 and 2016, respectively. He is currently a research assistant with the Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences. His research interests focus on deep learning, computer vision, and biometric recognition.

Jingtuo Liu received his B.S. and M.S. degrees in Electronics & Engineering from Tsinghua University in 2008 and 2011, respectively. He has been working on computer vision at Baidu, Inc. since 2011. He is now a chief architect of a computer vision team at Baidu, Inc. His research interests focus on face detection and recognition and optical character recognition.

Ran He received the B.E. and M.S. degrees in computer science from the Dalian University of Technology, in 2001 and 2004, respectively, and the Ph.D. degree in pattern recognition and intelligent systems from the Chinese Academy of Sciences in 2009. Since 2010, he has been a Full Professor with the National Laboratory of Pattern Recognition (NLPR). His research interests focus on information theoretic learning, pattern recognition, and computer vision. He serves as an Associate Editor of Neurocomputing (Elsevier) and serves on the program committee of several conferences.