Knowledge-Based Systems 253 (2022) 109499

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Scale adaption-guided human face detection

Cunying Ye [a], Xin Li [a], Shenqi Lai [a], Yaxiong Wang [a,*], Xueming Qian [a,b,**]

[a] School of Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, 710049, China
[b] Ministry of Education Key Laboratory for Intelligent Networks and Networks Security, China

ARTICLE INFO

Article history:
Received 29 April 2022
Received in revised form 12 July 2022
Accepted 18 July 2022
Available online 27 July 2022

Keywords:
Face detection
Anchor-free
Scale adaption
Context aware

ABSTRACT

Anchor-free based object detection has recently seen important progress benefiting from the advances in convolution neural networks. However, the detection performance for human faces is not so satisfactory. First of all, many existing anchor-free methods only focus on a certain scale of the feature map, such a mechanism often fails to perceive the important multi-scale context, resulting in a low recall rate of faces with large scale variations. To solve this problem, we propose to boost the face detection by adaptive learning to perceive the focal scale. To be specific, we design an online scale adaptation strategy to heuristically guide each layer detector to detect faces of different scales in multi-branch structures, which reduces outliers and improves recall rates. In additional, we also argue that the detection head with single convolution layer widely used in anchor-free methods is not robust enough to image context. Therefore, we augment the network by a context-aware detection module. The module dynamically generates different detectors for different input images based on their context to adapt to their image features, reducing the dependence on feature extraction ability of backbone network, and avoiding feature deviations in different scenes. Extensive experiments demonstrate that our method achieves significant performance gains compared to previous anchor-free methods and is comparable to the most advanced anchor-based face detection methods.

1. Introduction

Face detection is a fundamental task in computer vision. It serves as primary technique for various downstream vision applications such as face recognition [1–3], face alignment [4–6] and face retrieval [7,8]. Thanks to the well-annotated datasets [9–12] and the rapid development of deep learning, a series of excellent algorithms [13–23] have been proposed in the past decades, which could be categorized into two branches, i.e., anchor-based and anchor-free methods. Comparing to the anchor-based detection methods, anchor-free methods could achieve a faster inference but worse performance. Extensive efforts have been dedicated to mitigate the performance gap between the anchor-based and anchor-free methods. However, the results are still unsatisfactory due to some challenging factors such as scale variation and complexity of image context. Scale variation is a long-standing challenge for anchor-free face detection. To address this issue, some anchor-based methods [13–19] utilize multiple feature maps from different layers and detect faces at each feature layer in parallel. And then they densely place anchor boxes with large size to upper feature maps and anchor boxes with small size to lower feature maps. The success of these methods mainly stems from the support of extensive anchors. Thus, these solutions could not be directly transferred to anchor-free methods since there is no anchor available. Keypoint-based anchor-free object detection is another popular detection paradigm [24–26], which abandons the anchor and treat object as a combination of some points. However, many existing keypoint-based models detect objects of various scales from a single scale feature map, and are not robust to handle the problem of large scale variation. For example, the prediction bounding box of large object is too small due to the lack of receptive field, and the keypoint location of small object is not accurate. Therefore, these methods also cannot be directly used for face detection with large scale variation. Recently, Li et al. [27] propose a pyramid feature aggregation mechanism to enhance the model robustness for the face scales, which could achieve a faster inference but worse performance than most advanced face detection approaches.

The effectiveness of multi-scale has been validated by anchor-based methods in addressing the scale variations [13–19], but the situation becomes much more challenging when encountering the anchor-free cases. Unlike the anchor-based methods that could utilize the IoU of anchor box and ground-truth to build the corresponding between the detectors and the faces with different scales, no anchor box is available for anchor-free

[*] Corresponding author.
[**] Corresponding author at: School of Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, 710049, China.
E-mail addresses: wangyx15@stu.xjtu.edu.cn (Y. Wang), qianxm@mail.xjtu.edu.cn (X. Qian).

https://doi.org/10.1016/j.knosys.2022.109499
0950-7051/© 2022 Elsevier B.V. All rights reserved.

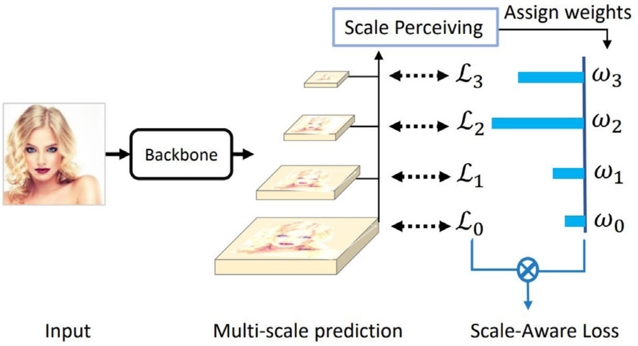

Fig. 1. The figure demonstrates the weight of training sample at different detection head under scale aware loss. The training sample will be assigned different loss weights to detectors of different levels by scale perceiving.

methods. To tackle this and successfully adopt the multi-scale mechanism into anchor-free community, we transform the original encoder–decoder structure in [24–26,28] into a multi-branch network suitable for face detection and propose a scale-aware loss to help the network perceive the scale difference, such that an explicit connection between face scale and detector could be built. The loss aims to make each detector focuses on samples near the reference scale and reduces the weights of samples far away. As shown in Fig. 1, the training sample will be assigned different loss weights to detectors of different levels by scale perceiving. For example, the scale of the input sample is closest to the reference scale of the second layer detection head, so we let the second layer detection head primarily detect the input sample by assigning it larger weight. Furthermore, to alleviate the sample imbalance of different scales, we design an data-scale-resampling method. This operation changes the distribution of samples with each scale in training dataset, making the number of all scales samples be balanced, and each detector could get enough samples for training. The scale-aware loss and data-scale-resampling method form our online scale adaption strategy, these two modules could help us build a scale-robust network.

In our practice, we also observed that the aspect ratios of faces are relatively diverse. Although aspect ratios of most faces are close to 1 : 1, according to the statistics of WIDER FACE [12] dataset, there are still a considerable number of face bounding boxes with an aspect ratio greater than 2 or less than 0.5. For these samples with singular aspect ratios, the common-used convolution with square receptive field could not well capture their key features. Consequently, the network often fails to detect the face regions. To address this, we design a shape-sensitive module (SSM). Our SSM combines standard convolutions and asymmetric convolutions, providing receptive fields with different shapes and enhancing the perceptual ability for faces with different shapes.

The contributions of this paper are summarized as follows:

- We propose an online scale adaption strategy that heuristically guides the detector on each layer to adapt faces with different scales in multi-branch structure. This strategy enhances the robustness of the network about the scale variation.
- In view of the context complexity of different images, we also propose a context-aware dynamical detection head. It dynamically generates different detectors according to the image content, reducing the dependence on the feature extraction ability of backbone network and preventing feature deviations between different scenes.
- We also propose a shape-sensitive module to improve the recall ratio of faces with singular aspect ratios. We evaluate the proposed method on popular face detection benchmarks FDDB [11], AFW [9], PASCAL face [10] and WIDER FACE [12] datasets. Extensive experimental results demonstrate the superiority of our proposed method with other state-of-the-art methods.

2. Related work

Face detection is a classic task in computer vision, and has been extensively studied over the past few decades. Early face detectors are based on sliding windows and hand-crafted features, while the modern methods are based on convolution neural networks. The CNN based face detection approaches can be roughly divided into two categories: anchor-based detector [13–17,19,29] and anchor-free detector [21,22]. These two stage methods [30, 31] use ROI Pooling operation [30] to extract a scale-invariant feature map for multi-scale detection, while some one-stage methods [21,22,24–26,32] also extract a single-scale feature map but lacks scale invariant. Recently detectors [13–17,19,29,33] adopt multi-scale feature maps for object detection. Considering that deep learning approaches have achieved the state-of-the-art results on all open datasets and far better than the traditional detection methods, here we mainly introduce the related deep learning methods.

2.1. Detect face based on anchors

Anchor-based detectors are the most popular and best performance methods for face detection. These methods use multi-scale anchor boxes at each cell of image to replace different sliding windows of traditional methods and rely on classifying anchor boxes to detect face. Anchor is first proposed by Faster R-CNN [31], and then SSD [33] extends anchor boxes to multi-scale feature maps. Since then, anchor-based methods [31,33,34] have achieved remarkable performance in the area of general object detection and anchor is rapidly used in the area of face detection. After several years of research, anchor-based methods [13–17, 19,29,35] have achieved great success in face detection task and won all champions in WIDER FACE Challenge [12] in a long time. Specifically, SSH [13] is designed to reduce inference time, which has small memory and scale invariance. It is a single-stage detector that classifies and locates the global information extracted by the convolutional layer. S3FD [14] can be regarded as an improvement based on SSD [33]. It improves the detection network by adding a prediction layer and sets more reasonable anchors by referring to the effective receptive field. A scale-compensated anchor matching strategy is adopted to increase the number of positive sample anchors to improve the face recall rate. In recent works, researchers focus on how to improve the recall rate and precision of face detection. SRN [17] selectively applies two-step classification and regression on different layers to reduce false positives and improve location accuracy simultaneously. Pyramidbox [16] fully exploits the context information to provide extra supervision for small faces.

2.2. Anchor-free face detection

Other methods attempt to directly predict the bounding boxes without pre-defined anchor boxes, called anchor-free detectors. Previous works [21,22] directly regress a 4-D vector at every pixel of feature map to locate faces by a fully convolutional network. AFN [36] leverages the local and global contextual information fusion to improve recall-rate of anchor free method. SAFD [37] uses the dilated convolution layers and attention mechanism to select the informative features that can accommodate to different scales. Feng et al. [38] propose a novel network with anchor-free detection and improve the performance in dense object detection by an altered feature enhancement module. However, as suffering from imbalance problem of positive and negative samples and lack of multi-scale information, these methods cannot achieve very promising results on the large-scale WIDER FACE Challenge [12].

Recently, there are some methods [24–26] that adopt keypoint estimation to locate objects. These methods decompose object as a combination of some points, and employ a huge neural network, such as Hourglass [39], to generate a high-resolution feature map to locate keypoints of objects. Instead of using embedding of each corner to achieve keypoint matching like CornerNet, Ma et al. [40] predict a matching degree score for each predicted bounding box formed by corners. Yang et al. [41] adopt two simple modules, ship detector and center detector based on the extracted features by the feature extraction module and the central information of ships to generate and improve the detection results. Unfortunately, these methods still lack multi-level detection and cannot handle with the imbalance problem among objects with different sizes. Furthermore, one anchor free method [27] proposes a pyramid feature aggregation module, which aggregates multi-layer features to enhance the modeling ability of the model to faces with different scales. The detection speed can be increased by two to three times. However, there is a small gap between the accuracy and the most advanced face detection methods.

2.3. Multi-scale enhanced object detection

Current CNN-based detectors extract feature maps from backbone network and then apply detection head on these feature maps to parse final detection results. Some detectors extract a single scale feature map for detection, such as [21,22,24–26,30–32]. These two stage methods [30,31] use ROI Pooling operation [30] to extract a scale-invariant feature map for multi-scale detection. They provide excellent performance on most common objects, while the recall rate of small objects is not high. One of the most important reasons might be that small objects are more difficult to match anchors and there are no enough positive samples of small objects for training. Some one-stage methods [21,22,24–26,32] also extract a single-scale feature map. However, as there is no ROI pooling operation in one stage methods, these features are lack of scale invariant, thus they perform not well on small objects and very large objects.

Other recently detectors [13–17,19,29,33,42–44] adopt multi-scale feature maps for object detection. They handle objects with different scales at different level feature maps and achieve great performance improvement. However, the performance of multi-scale detectors is highly related to the design of anchors. If one object does not match pre-defined anchor boxes, the recall rate will be not high. FaceBoxes [19] proposes a rule for anchor design that makes anchors with different scales have the same density on the input image, thus the recall rate of tiny faces has been greatly improved. S3FD [14] proposes a scale compensation anchor matching strategy to ensure that faces with each scale could match enough anchors. RetinaFace [29] places over 100,000 anchor boxes on multi-scale feature maps to improve face detection performance. Wu et al. [45] propose a novel object detection framework that integrates multiple channel feature extraction, feature learning, fast image pyramid matching and boosting strategies. In our paper, we propose a scale adaptive loss, which adjusts the loss weight of samples in different detectors according to the scale and heuristically guides each detector to learn an optimal detection range, rather than directly assign samples to detectors of different levels.

3. Preliminary

In this paper, we extend keypoint-based anchor free detection methods [24,26,27,46]. At first, we briefly review the anchor free detector. Similar to [27], we decompose face detection into two tasks: center localization and scale prediction. First, an input image goes forward through the stacked convolutional layers to form feature maps. Based on the feature map, anchor free detector separately generates two heatmaps: Center map and Scale map, as shown in Fig. 3. The center map shows where the face may exist, and the scale map is used to predict the scale of face at every position.

For center localization task, we formulate it as a binary classification task (center or non-center) at pixel level. The box center is treated as the positive target while others are treated as negative samples, and we employ 2D Gaussian mask that reduce the penalty for negative samples near the center point to deal with the imbalance of positive and negative samples. For scale prediction task, we predict a 4-D vector that is the distances from its location to the four sides of the bounding box at each position. The training loss is composed of the loss of center point estimation and the loss scale regression:

$$L_{det} = L_{center} + \lambda \cdot L_{scale} \tag{1}$$

where L_{center} and L_{scale} are the loss of center estimation and scale regression respectively. λ is a hyper-parameter to balance loss between center point estimation task and scale prediction task. More details about the anchor-free detector and the training procedure can refer to [27].

4. Proposed method

Our overall architecture is exhibited in Fig. 2, given an input face, we first adopt our data-scale-resampling strategy and select an image with certain scale based on a pre-defined scale pool. Then, the image is fed into the network to produce the multi-scale features. In the following, the features with different scales are first up-sampled and pass through our shape-sensitive module (SSM) and the context-aware detection module dynamically generates the detectors according to the input feature content, predicts the center point position and the corresponding scale information of the face to obtain the detection bounding box. Finally, the proposed scale-aware loss is employed to train the network. In this section, we will introduce our online scale adaption strategy, shape-sensitive module and context-aware dynamical detection module, respectively.

4.1. Online scale adaption strategy

To apply the multi-branch parallel detection manner in the anchor-free based framework, how to make different detectors adapt to different scales is a key problem to be addressed. The traditional anchor-based methods assign training samples to different levels by the IoU between ground-truth and anchor boxes, thus the detector of each layer only needs to focus on the objects that match anchor boxes of this layer. However, because there is no anchor as a reference in anchor-free methods, we cannot directly bind training samples with different scales to different detectors in training process.

To successfully integrate the multi-scale mechanism into keypoint anchor free paradigm, we propose an online scale adaption strategy which comprises the scale aware loss and data-scale-resampling, these two modules work together to mitigate the problem of scale variation. The scale aware loss is designed to heuristically guide detectors at each layer to learn a reasonable detection range, while the data-scale-resampling strategy is responsible to change the distribution of training samples with different scales and make the samples of each detectors more balanced.

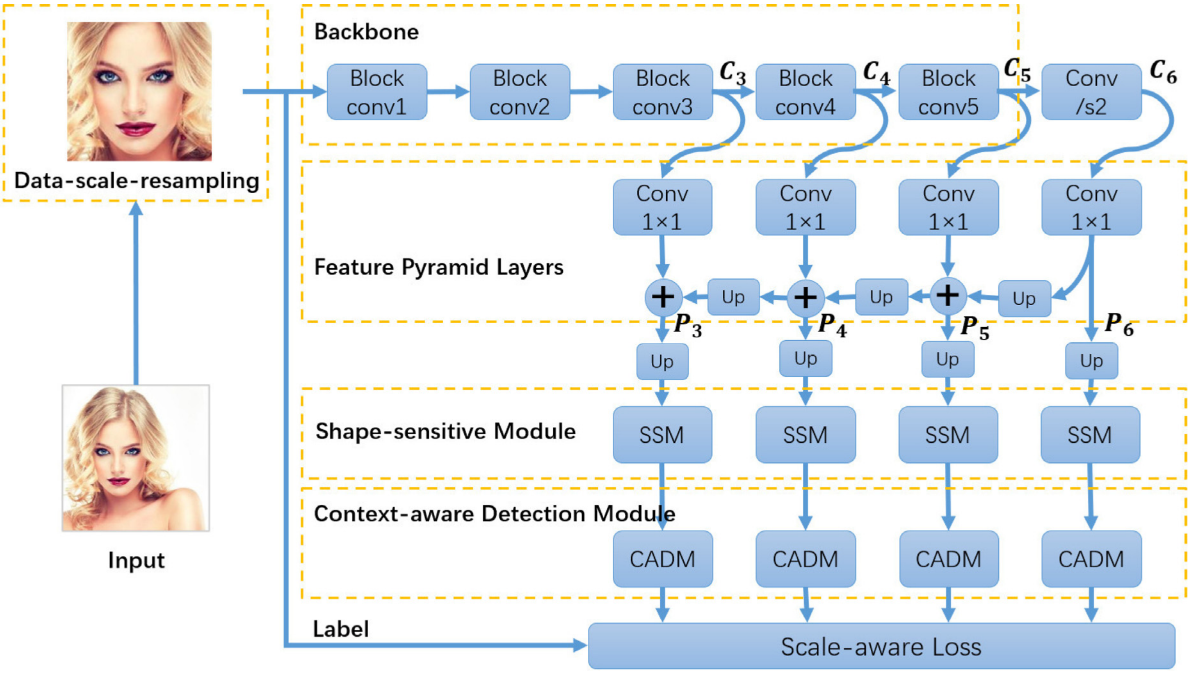

Fig. 2. Network architecture used in our experiment, *Up* is a bilinear upsampling, *SSM* is shape-sensitive module and *CADM* is context-aware detection module.

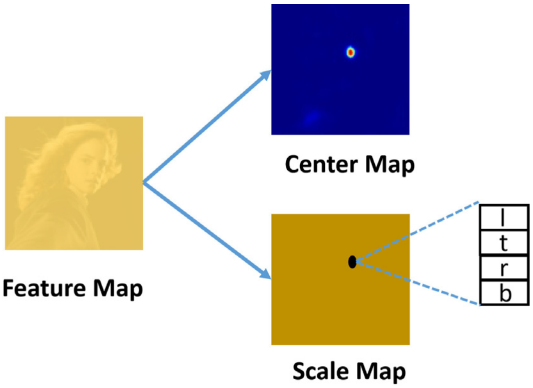

Fig. 3. An illustration of anchor-free detector, center map and scale map are generated on feature map.

4.1.1. Scale-aware loss

Our scale-aware loss targets on guiding each detector to pay more attention to the samples whose scales are near the reference scale, this is achieved by assigning a scale-adaption weight for the loss of respective detector.

As shown in Fig. 2, our network employs a multi-branch architecture and outputs multi-scale prediction by different detectors. We first estimate a weight for each detector based on the face scale, which is used to adaptively accumulate the losses of all detectors, such that the overall loss could be obtained. In particular, we formulate the total loss of a face e in all branch detectors as:

$$L(e) = \sum_{i=0}^{k} p^i \cdot \left(L^i_{center}(e) + \lambda \cdot L^i_{scale}(e)\right) \tag{2}$$

where superscript i represents different detectors, k represents the number of all detector heads, L_{center}, and L_{scale} are the loss of center estimation and scale regression respectively, which is introduced in Section 3, λ is a hyper-parameter to balance loss between center point estimation task and scale prediction task, p^i is the weight of the loss in the ith branch detector.

As shown in Eq. (2), the key of our scale-aware loss lies on how to determine the weights p^i. Intuitively, we attempt to guide each detector to focus more on samples near the reference scale and reduce the weight of the sample that is far away, so the form of the Gaussian function is more in line with our requirements. Specifically, p^i is a parameter related to the scale s of training sample and the reference scale ref_i. According to previous research [14], the effective receptive field of each layer in CNN model is about 4 times of its total stride. Thus, each feature map is more suitable for detecting these faces whose sizes are similar to 4 times of its stride. So we set a reference detection scale ref_i for each detector with $4 \times stride_i$. When the scale s of the training sample is further away from the reference scale ref_i, the weight p^i of the sample in this branch is smaller.

The weight p^i is calculated by:

$$\begin{aligned} ds_i &= max(\frac{s}{ref_i}, \frac{ref_i}{s}) \\ p^i &= exp\left(\frac{-ds_i^2}{\tau}\right) \end{aligned} \tag{3}$$

where ref_i is the reference detection scale that we set according to the stride in the ith detector. τ is a hyper-parameter, which controls the decay speed of weight. For a training sample, we take the diagonal length of bounding box as its scale s:

$$s = \sqrt{w * h}. \tag{4}$$

where w and h are the corresponding width and height of the bounding box.

4.1.2. Data-scale resampling

The distribution of training samples with different scales is imbalanced. The number of samples with some scales is very large while the other is very small, which causes the network bias to the scales with enough images and get barren performance on the insufficient scales. To remedy this issue, we design a data-scale-resampling strategy to change the distribution of faces with different scales in training dataset, which makes faces with every scale more balance and make sure that detectors at each level can get enough samples for training.

The main idea is to randomly select a face from input image and then rescale the selected face to a randomly selected scale. In this way, we can get a relatively even samples with all scales. Supposing $\mathcal{B}$ is the set of bounding boxes and faces, and $\mathcal{R} = \{ref_i\}_{i=0}^{k}$ is the set of reference scales. Select a face bounding box randomly from $\mathcal{B}$ and calculate the scale s of the selected face bounding box. Then select a reference scale randomly from $\mathcal{R}$ and in the fourth step, select a target scale around the reference scale. Calculate the scaling factor r^*. Finally, resize the input image by the factor r^* and crop a region around the selected face in the scaled image as a training image. The detailed steps are described in the following algorithm 1.

Algorithm 1: Data-scale-resampling.

```
# B: the set of bounding boxes and faces
# R: the set of reference scales

B, R = input(B, R) # input the B and R
# randomly select a bounding box and its face
face, box = random.choice(B)
# calculate the scale of the selected face
s_face = sqrt(w * h)
# randomly select a reference scale
ref = random.choice(R)
# select an object scale around the reference scale
s_target = random.uniform(ref / 2, ref * 2)
r = s_target / s_face # scaling factor
# resize and crop a region around the selected face as a
    training sample
scaled_face = crop(r, face)
```

Tang et al. [16] also develop a data sampling strategy to remedy the scale imbalance problem. However, the implementation is consideratly different with our data-scale resampling. The authors pre-define a anchor scale pool, for a training image, first pick a closest anchor scale to the image, then reshape a random face in this image to a random smaller anchor scale. While in our Data-Scale Resampling, instead of giving a fixed scale pool beforehand, we introduce a reference set, from which we pick a scale to serve as a reference for the image scale and select a target scale around the reference scale.

4.2. Shape sensitive module

The shape of human face is different in scale and aspect ratio. As reported in [13,16], increasing the receptive field by placing convolution kernels with different sizes in parallel upon extracted feature maps can improve face detection accuracy. Indeed, the combination of convolution kernels with different sizes not only enlarges the receptive field, but also diversifies the receptive field of detection layers. The diversity of receptive fields increases the ability of model to capture faces with different sizes. However, the receptive fields of most networks are square, which will affect the detection of faces with different aspect ratios.

To enable the network to effectively perceive the faces with diverse size and aspect ratio, we design a Shape Sensitive Module(SSM) to enhance the expression ability of feature maps for faces with different shapes. Our SSM is a multi-branch convolutional block. It can be divided into two parts: the cross-perceiving part uses asymmetric convolutional layers to provide rectangular receptive fields to deal with these faces with different aspect ratios, and the square-perceiving part uses standard convolution and residual connection to provide square receptive fields with different scales to enhance the modeling ability of multi-scale faces.

In particular, the cross-perceiving part comprises four convolutions with receptive fields 3 × 1, 1 × 3, 5 × 1 and 1 × 5, while the square-perceiving part contains two shared convolutions which provide two receptive fields, 3 × 3 and 5 × 5 . Fig. 5 illustrates the mechanism of SSM. The color rectangle of each layer is the receptive field of one correspondent convolutional layer. Suppose the receptive field of feature is the top small square in Fig. 5, and the different rectangles in the middle layer are the receptive fields corresponding to each branch of the SSM module. The bottom layer is the final enhanced receptive field. The specific process is as follows: firstly input a feature map, divide the feature map into several groups in the channel dimension. Next input them into the cross-perceiving part and the square-perceiving part respectively for different convolutions. And then concat the two part to obtain the enhanced feature maps.

To be specific, the cross-perceiving part can be formulated as follows:

$$\hat{x}_{cp} = Conv_{3\times1}(x) + Conv_{1\times3}(x) + Conv_{5\times1}(x) + Conv_{1\times5}(x) \tag{5}$$

where x is the input original feature and $\hat{x}_{cp}$ is the enhanced feature. This enhanced feature have receptive fields with multiple shapes, such as 3 × 1, 1 × 3, 5 × 1 and 1 × 5.

The square-perceiving part can be formulated as:

$$\hat{x}_{sp} = Conv_{3\times3}(x) + Conv_{3\times3}(Conv_{3\times3}(x)) \tag{6}$$

where the $Conv_{3\times3}$ is shared. These two cascaded convolution layers provide 5 × 5 receptive fields. Thus, the square-perceiving part has two receptive fields, 3 × 3 and 5 × 5 .

Finally, we also provide a shortcut connection as an access to the flow of original information, which is equivalent to a 1 × 1 receptive field.

4.3. Context aware detection module

The CNN-based face detectors often extract several feature maps with different resolutions, and then use a detection head on these feature maps to parse out detection results (confidence and bounding boxes). The detection head usually consists of several convolutional layers. The drawback of the general convolution detection head is that its parameters will be fixed after finishing training and it lacks self-adaptability to image content. In order to obtain correct detection results, the fixed detection head requires that input features come from the same domain. It means that general detection head has a high requirement on the feature extraction capability of backbone network. However, for two images with large difference in context (such as one from normal illumination scene and the other from overexposure scene), the feature extracted by backbone always has a bias, which will affect detection results. A naive way is to train a detector for each scene(such as dark, overexposure or normal) separately. Unfortunately, explicit scene classification requires scene annotation on a large amount of data, which is very expensive and time consuming. Moreover, due to cognitive limitations of the annotators, various scenes cannot be well classified.

Hong et al. [47] develop a general multimodal deep learning framework to solve the problem of deep network performance degradation caused by classification tasks based on complex scenarios. And similarly, we propose a context aware detection module to solve this problem. It dynamically generates a detector according to the content of input image, as shown in Fig. 6, so that the generated detectors can be adapted to the features.

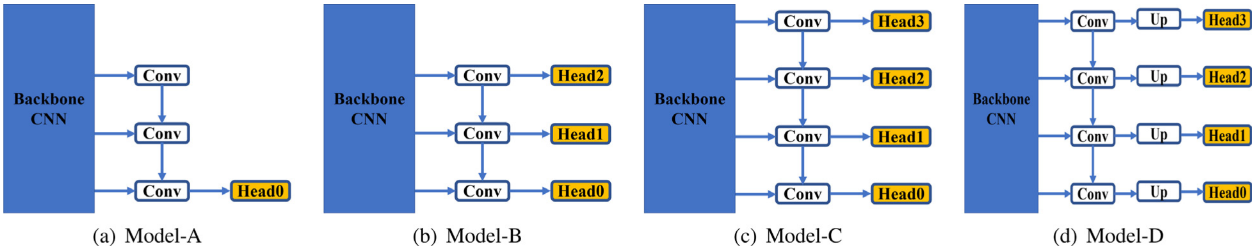

Fig. 4. Evolution of multi-branch model structure.

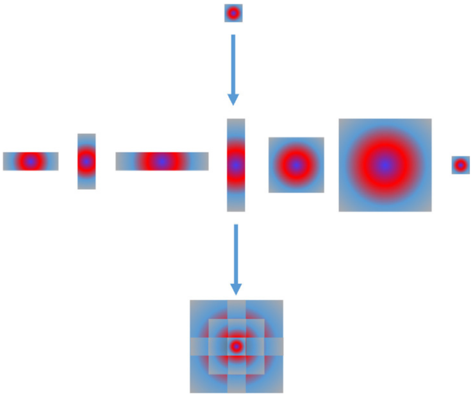

Fig. 5. The schematic diagram of receptive field enhancement.

This dynamic detector no longer relies too much on the feature extraction ability of backbone network and prevents the feature bias between different scenes.

Give a series of parallel convolution kernels, which can be considered as multiple parallel templates, and calculate the weight of each convolution kernel by *Gate* network according to the input feature. These templates are weighted to produce a new detection module, that is, the context-aware detection module is obtained. Firstly, we will introduce how to generate the weighted template T dynamically. Suppose $f_0, f_1, f_2 \ldots, f_{k-1}$ is a series of parallel convolution kernels. Then P_i is the input feature map and T_{P_i} is dynamic weighted template generated according to P_i.

We parameterize the dynamic weighted template as a linear combination of k templates, and it is defined as:

$$T_{P_i}(P_i) = w_0^i \cdot f_0 + w_1^i \cdot f_1 + \cdots + w_{k-1}^i \cdot f_{k-1} \tag{7}$$

where each w_k is an instance-dependent scalar weight computed using a simple *Gate* network. k represents the number of parallel convolution kernels and f_k represents the kth parallel convolution kernel. For different features, the weights calculated by *Gate* are different. Thus, a dynamic converter is generated by changing the weight of each template. We hope that the *Gate* network is computationally efficient, and can establish the relationship between input features and templates, so we design a simple network composed of global average pooling, fully connected layer and sigmoid activation as the *Gate* function. Suppose the input feature is P_i, firstly, the global average pooling(GAP) of P_i is carried out, and then the pooling result is passed through a fully connected layer(FC). Finally, sigmoid activation is performed to obtain w_k^i. The process of weight generation can be formulated as:

$$[w_0^i, w_1^i, w_2^i \ldots, w_{k-1}^i] = Sigmoid(FC(GAP(P_i))) \tag{8}$$

The detection process of our proposed context aware detection module can be formulated as:

$$Dets = \phi(T_{P_i}(P_i)) \tag{9}$$

where *Dets* is final detection results, ϕ represents a convolution operation.

Since different convolutional kernels collect contexts from different views, Eq. (7)–(8) actually seeks to form an adaptive context generation, where the adaptive weights are generated based on the input feature, with the softmax operation, we could endue the network with the ability of dynamic context selection, such that our network could be capable of picking the most valuable context to benefit the followed detection task.

5. Experiments

In this section, we evaluate our method on four common face detection datasets, FDDB [11], WIDER FACE [12], AFW [9] and PASCAL Face [10]. Following standard practice, all models are trained on the WIDER FACE dataset while other datasets are only used to evaluate the final performance. To show the effectiveness of the proposed method, comprehensive ablation studies and discussions are given.

5.1. Datasets

5.1.1. FDDB

The images of FDDB are collected from unconstrained natural scenes. It has 2845 images with 5171 annotated faces. These images have a wide range of difficulties, such as low images resolutions, make-ups, occlusions.

5.1.2. WIDER FACE

It is the most challenging public face detection dataset, images of which has dramatic variability in scale, pose and occlusion. This dataset contains 32,203 images with 393,703 labeled faces, and those images are split into training (40%), validation (10%) and testing (50%) set. For testing and validation sets, the images are divided into three levels (Easy, Medium, Hard subset) according to the difficulties of detection. Ablation studies are performed on the validation set.

5.1.3. AFW dataset

The images of AFW dataset are collected from Flickr. It has 205 high-resolution images with 473 annotated faces.

5.1.4. PASCAL face dataset

The images of this dataset are selected from PASCAL VOC dataset. It has 851 images with 1335 annotated faces.

5.2. Experimental setup

5.2.1. Data augmentation

To make the model more robust to input face sizes and prevent over-fitting, random crop data augmentation strategy is adopted. More specifically, we random crop a square patch for each training image with a random size between [0.3,1] of the original image's short edge. Then we rescale this patch to 640 × 640. Besides random crop, random horizontal flip and photometric color distortion [14] are also employed.

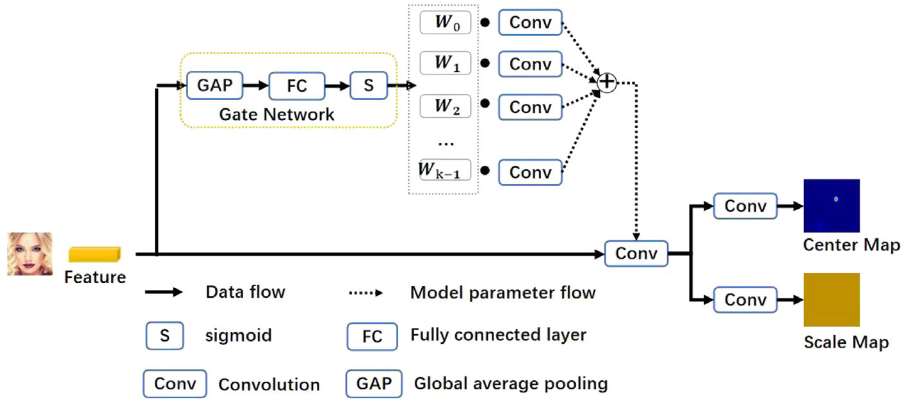

Fig. 6. The structure of context-aware dynamical detector. *GAP* is global average pooling, *FC* is fully connected layer, *S* is sigmoid function, *Conv* is convolution layer.

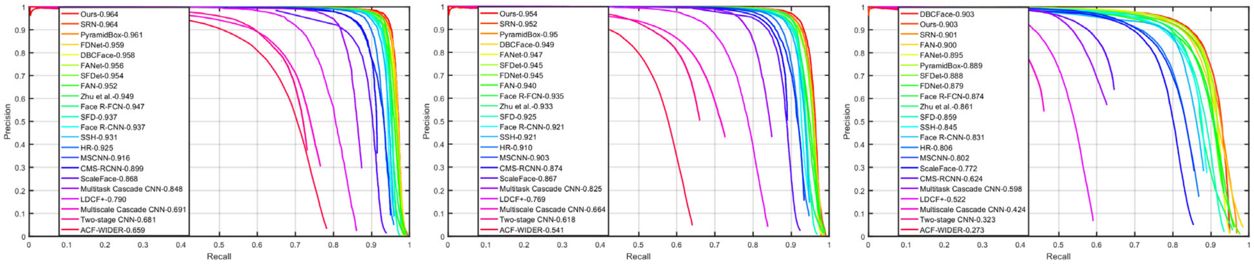

Fig. 7. PR curves and AP on WIDER FACE dataset. From top to bottom are the results on the three data subsets of easy, medium and hard.

5.2.2. Training & testing details

We use ResNet-50 [48] pre-trained on ImageNet [49] as the backbone for experiments. All model are trained by Adam optimizer with the batch size of 24. The learning rate η is set to 1.5×10^{-4} for the first 100 epoch, and divided by 10 at 100 and 120 epoch. For the datasets covered in this article, WIDER FACE and the other three datasets will perform different test strategies. For WIDER FACE dataset, we will follow the standard strategy [16,17] that multi-scale testing and box voting are used to produce 750 best scoring results.

5.2.3. Baseline

We implement a fully convolution anchor-free face detector as baseline. Specifically, we construct a Feature Pyramid Network(FPN) [18] and attend a general anchor-free detection head on $P3$ level, which has 1/8 resolution of input image. This baseline can be called single scale anchor-free face detector, in which it predicts face center point and regresses face size directly from image feature map (P3). Unless otherwise stated, all experiments are carried out on WIDER FACE dataset.

5.3. Model architecture design

We design four kinds of network structures, as shown in Fig. 4. Model-A is our baseline and represents a series of recent keypoint-based detection methods. This series of methods detect all scale objects on a fixed size feature map. Model-B is a standard multiple scale detector based on FPN. Model-C adds a new scale detector to model-B. Model-D increases the size of feature map by adding up-sampling structure on each feature map of Model-C. For these three multi-branch structures, Model-B, Model-C and Model-D, we use hard interval division and online scale adaptive strategies to adjust the detection range of detectors on different layers. How to produce reference scale has been described in Section 4.1.1. In hard interval division strategy, we empirically set the threshold of scale between two adjacent detectors as $ref_i + \frac{ref_{i+1} - ref_i}{2}$.

5.4. Ablation study

5.4.1. The effect of online scale adaption

As there is no anchor as reference in anchor-free methods, it is vital for multi-branch structure training that how to determine the detection range of detectors at different level. To demonstrate the effectiveness of online scale adaption strategy, we conduct a series of controlled experiments and results are shown in Table 1. The baseline is a single branch structure that detects faces of all scales on a fixed feature map.

Based on the results in Table 1, we can see that the interval division strategy performs worse on Model-B, Model-C and Model-D, and even is worse than baseline. This hard assignment method makes it difficult for the model to deal with critical samples near the threshold. It leads to a low recall rate for these faces whose sizes are close to threshold. Thus, it cannot obtain the performance gain from multi branch structure. The results in Table 1 show that our online scale adaption strategy greatly outperforms interval division strategy in these all three multi-branch models. The comparison between the fourth and the fifth lines in Table 1 indicates that our online scale adaption strategy effectively improves the performance, especially for big faces. The AP is increased by 5.01, 2.32, 2.12 on easy, medium and hard subset, respectively. The increase mainly comes from higher recall rate of faces with various scales. Those faces whose sizes are close to threshold may be detected in everyone between two adjacent detectors because this scale samples are considered in all level detectors and only the importance is different. This multi-level detection ensures high recall rate. And it can be observed in Table 1 that more branches can achieve a better performance. The comparison between the first and third lines in Table 1 shows the AP is increased by 1.87, 0.38 and 0.74. Also can be seen,

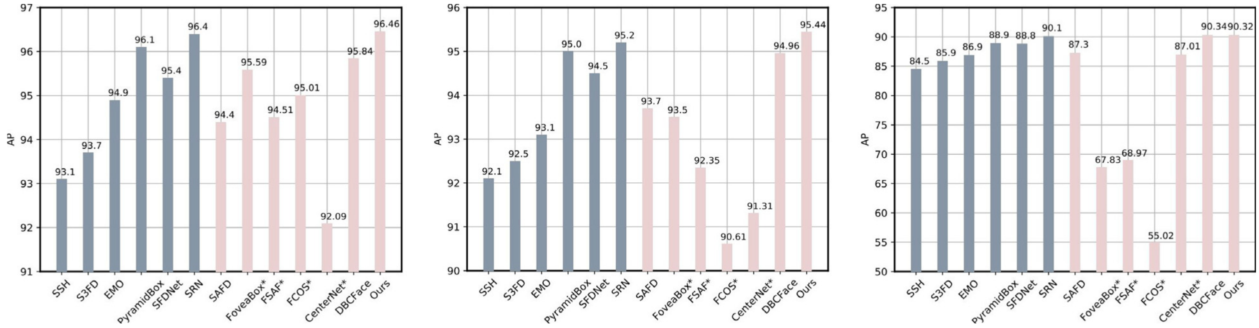

Fig. 8. Performance results on the validation set of WIDER FACE. From left to right are the results on the three data subsets of easy, medium and hard. And for each of these figures, the part with the same color on the left are the anchor-based methods, and the one on the right are the anchor-free methods. ⋆ means that this method is originally designed for general object detection, and we reimplement it for face detection.

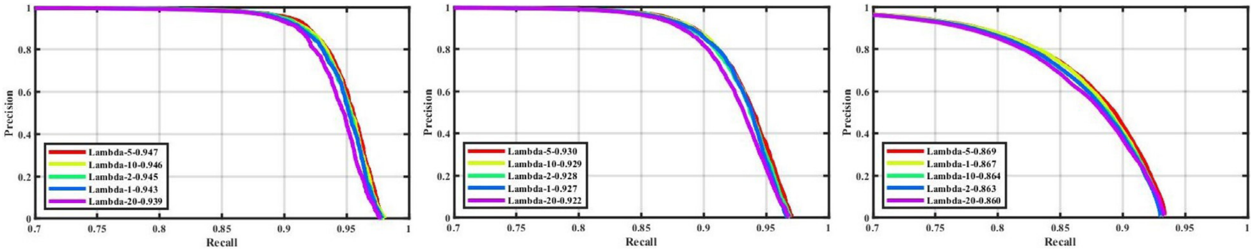

Fig. 9. The effect of λ on WIDER FACE dataset. The performance is evaluated under the model-D structure of ResNet-50 backbone.

Table 1
Effect of various model structure designs and different detection scales allocation strategy on WIDER FACE dataset.

Model	Backbone	Interval division	Online scale adaptive	Multi-scale test	Easy	Medium	Hard
Model-A(baseline)	MobilenetV1 0.25x				86.57	85.00	70.41
Model-B	MobilenetV1 0.25x	✓			83.72	83.13	69.42
Model-B	MobilenetV1 0.25x		✓		88.44	85.38	71.15
Model-C	MobilenetV1 0.25x	✓			84.10	83.78	69.56
Model-C	MobilenetV1 0.25x		✓		89.11	86.1	71.68
Model-D	MobilenetV1 0.25x		✓		90.08	86.57	71.83
Model-D	MobilenetV1 0.25x		✓	✓	92.91	90.76	82.14
Model-D	ResNet-50		✓	✓	95.92	94.81	89.86

Table 2
Effectiveness of our proposed shape sensitive module on WIDER FACE dataset. The performance is evaluated under the model-D structure of ResNet-50 backbone.

Model	Easy	Medium	Hard
Model-D	95.92	94.81	89.86
Model-D + SSH	96.12	95.18	89.87
Model-D + SSM	**96.31**	**95.23**	**90.18**

the performance of four branches (Model-C) is better than three branch (Model-B). The multiple branch detection structure is an effective method to deal with the scale variation of objects.

Moreover, to reduce the discretization error caused by down sampling, we add a up-sampling structure on each feature map to generate finer feature map, as shown in Fig. 4(d). The comparison between Model-C and Model-D shows that the AP is increased by 0.97, 0.47 and 0.15 on easy, medium and hard subset, respectively. The results indicate that the upsampling operation has a better improvement for large targets. The reason might be that large faces are usually detected on high-level feature maps, and the downsampling rate of high-level feature map is relatively high. Higher downsampling rate will lead to higher discretization error. When we use the upsampling module to extract more fine features, these discretization errors will be reduced.

5.4.2. The effectiveness of shape sensitive module

The SSM uses asymmetric convolution to capture faces with extreme shapes. In previous work, SSH [13] uses filters of different sizes to capture the context of faces. However, the bounding boxes of faces are not always in the shape of a square. The square receptive fields may affect the detection of faces with different aspect shapes. We add a set of parallel asymmetric convolutions to capture faces with different shapes. Comparing the results between first and third lines in Table 2, we notice that SSM significantly improves the AP by 0.39, 0.42 and 0.32 on easy, medium and hard subset. Compared with SSH, our method also has better performance because SSH only increases receptive field while our SSM provides diverse rectangular receptive fields that increase the performance of faces with different aspect ratios.

5.4.3. The number of convolution kernels in context-aware dynamical detector

We use a set of parallel convolution kernels to dynamically generate different detectors according to the image content. The number of convolution kernels will affect the detection results. We compare the detection performance under different number

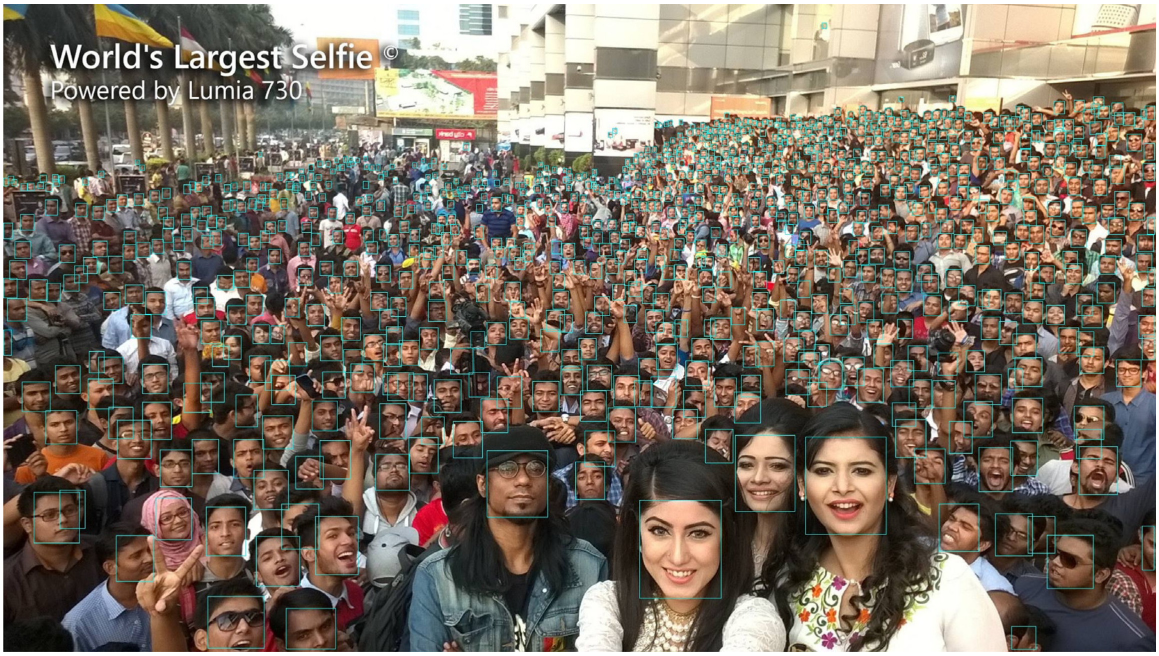

Fig. 10. Impressive qualitative result. Our model finds over 900 faces out of the reported 1000 faces.

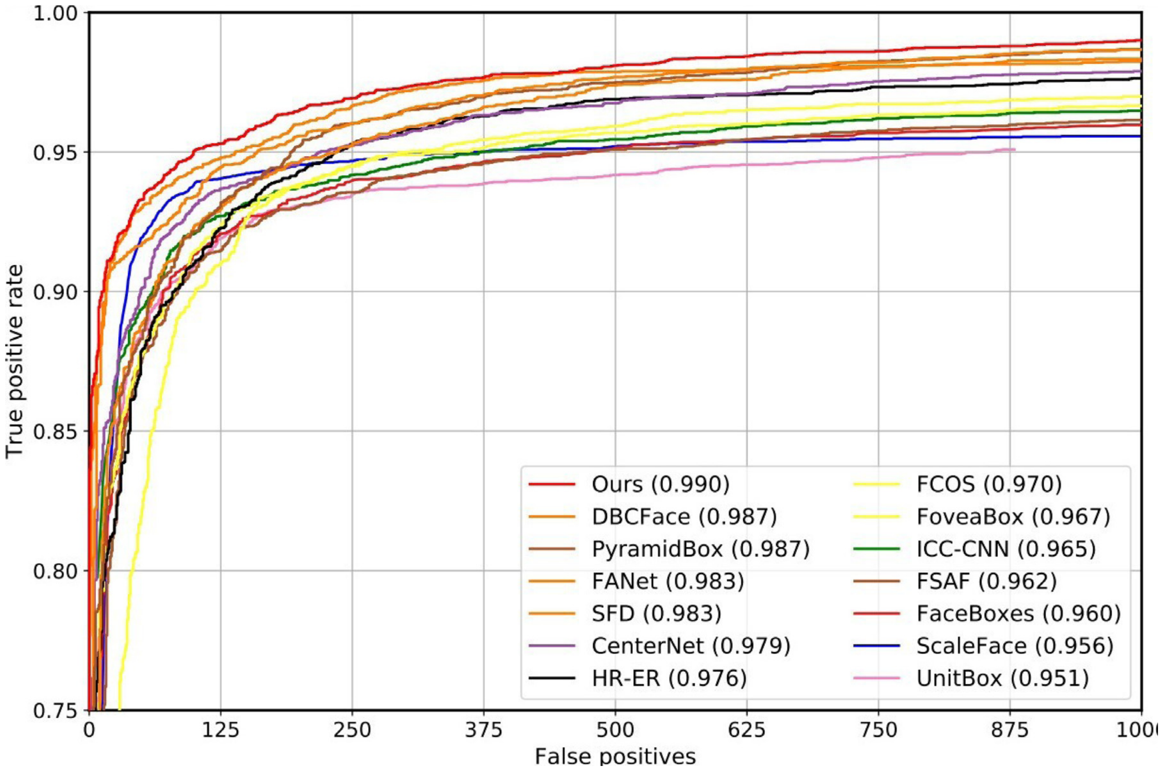

Fig. 11. The ROC curve and AUC on FDDB dataset by using the "discrete score" evaluation criteria.

of kernels, as results shown in Table 3. When the number of convolution kernels becomes one, the context-aware detection module will degenerate to general detection head. As the number of kernels grows, the performance first improves and then tends to be stable. Under such a high baseline, the performance are also improved by 0.15, 0.21,0.14 on easy, medium and hard subset when the number of kernels increase from 1 to 4.

5.4.4. *The effect of hyper-parameter λ in Eq. (2)*

We give the experimental results under model-D structure of ResNet-50 backbone, λ values are 1, 2, 5, 10 and 20 respectively. Fig. 9 is the result on easy, medium and hard subsets of WIDER FACE dataset. From Fig. 9, the model with λ=5 achieves the best performance, however λ does not have a remarkable effect on the final performance as shown in Fig. 9, revealing our method is

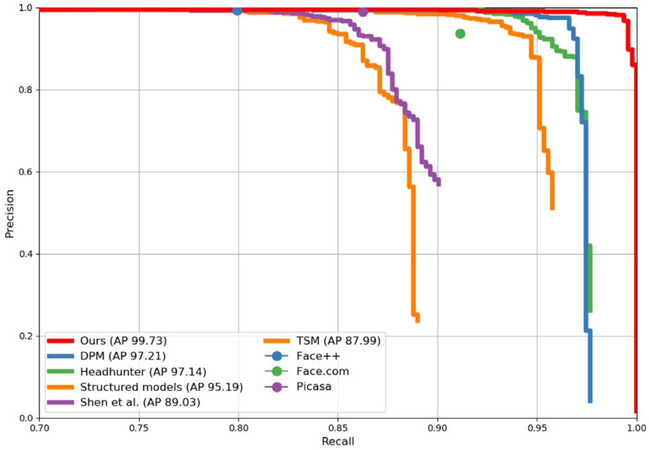

Fig. 12. Precision–recall curves on the AFW dataset.

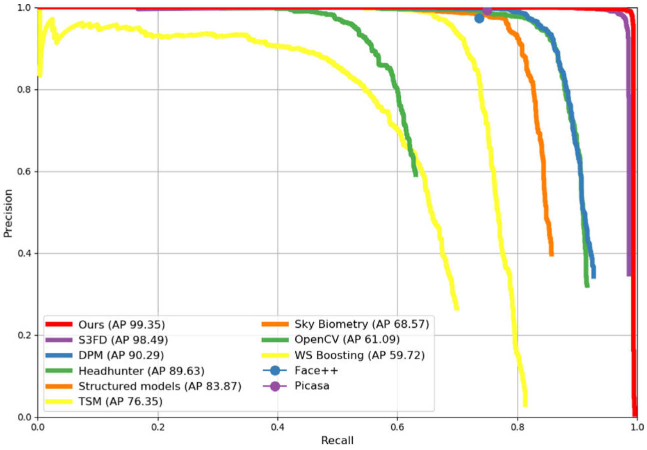

Fig. 13. Precision–recall curves on the PASCAL face dataset.

Table 3
The effect of convolution kernel number on WIDER FACE dataset. The performance is evaluated under the model-D structure of ResNet-50 backbone.

Number	Easy	Medium	Hard
1 (Model-D +SSM)	96.31	95.23	90.18
2	96.39	95.33	90.23
4	**96.46**	**95.44**	**90.32**
6	95.45	95.42	90.31
8	96.46	95.40	90.29

Table 4
Detection time with respect to different input sizes.

Method	640 × 480	1280 × 720	1920 × 1080
SRN [17]	88.45 ms	158.11 ms	309.95 ms
PyramidBox [16]	61.72 ms	166.21 ms	410.22 ms
DBCFace [27]	29.11 ms	63.73 ms	141.46 ms
Ours	40 ms	80.81 ms	139.76 ms

robust to this hyper-parameter. It is worth noting that due to the recent shortage of Our GPU resources, we set batch size as 8 in this experiment, which is much smaller than 24 set in our other experiments, so the experimental results are a little bit worse than those in our paper.

5.5. Evaluation on benchmark

We evaluate our model on the common face detection benchmarks, including WIDER FACE and FDDB. We find that our model achieves comparable results against other state-of-the-art methods on these two datasets, i.e. 96.46, 95.44 and 90.32 on easy, medium and hard subset of WIDER FACE, 99.0 on FDDB dataset. We also show a qualitative result of World Largest Selfie in Fig. 10. Our model can successfully detect over 900 faces out of 1000 faces reported.

5.5.1. WIDER FACE

We compare our method with the state-of-the-art face detection methods [13,14,16,17,27,36,37,50] and the state-of-the-art object detection methods [46,51–53] on WIDER FACE val subsets. For a more comprehensive comparison, we evaluate the performance of some typical the state-of-the-art anchor-free methods [46,51–53] on face detection datasets. The results are shown in Figs. 7 and 8. We can see that our method achieves 96.46, 95.44 and 90.32 on the three subsets, and outperforms other anchor-free methods by a large margin. Comparing with state-of-the-art anchor-based face detection methods, our method also achieves competitive performance.

5.5.2. FDDB

We evaluate our proposed method on the FDDB dataset and compare it with other state-of-the-art methods [14,16,19,22,27,54–61]. The discrete ROC curves are shown in Fig. 11. We can see that our method achieves the best performance over other state-of-the-art methods in terms of ROC curve.

5.5.3. AFW dataset and PASCAL face dataset

We evaluate our method on the AFW datast and PASCAL face dataset and compare the proposed method with some well-known works and three commercial face detectors (Face.com, Face++ and Picasa). Due to these two datasets are a little old, we only use to verify the generalization of our model. The precision–recall curves on AFW and PASCAL face are shown in Figs. 12 and 13, respectively. The average precision (AP) of our method on AFW dataset is 99.73 and on PASCAL face dataset is 99.35. The results show that our method is superior to the others and AP tends to saturate on both datasets, indicating that our method has good generalization.

5.6. Inference time

We analyze the running speed of our method on a single NVIDIA GTX 2080Ti. The running speed is the real detection time (including forward time and post-process time). We use batch-size 1 and a few common resolutions for testing. We average the time on WIDER FACE validation set to obtain reliable results. For comparison, we test two famous state-of-the-art anchor based methods PyramidBox [16] and SRN [17] and one anchor-free method DBCFace [27] under the same configurations. The final results are presented in Table 4. As can be seen, our method can achieve higher speed than two anchor-based methods. The anchor-free method DBCFace has faster speed than ours, but we can achieve better accuracy. Overall, our model achieves a good trade-off between performance and efficiency.

5.7. Computational complexity

We use fvcore[1] to compute the parameters of our model and FLOPs. Fvcore is a light-weight core library that provides the most common and essential functionality shared in various

[1] The open source of fvcore locates at: https://github.com/facebookresearch/fvcore

Table 5
The computational complexity of our model and DBCFace.

Model	Input	GFLOPs	Params(M)
DBCFace [27]	(1, 3, 256, 256)	16.76	37.6
DBCFace [27]	(1, 3, 512, 512)	67.05	37.6
DBCFace [27]	(8, 3, 256, 256)	134.1	37.6
Ours	(1, 3, 256, 256)	18.19	39.5
Ours	(1, 3, 512, 512)	72.49	39.5
Ours	(8, 3, 256, 256)	145.54	39.5

computer vision frameworks, including compute the parameters of the model and FLOPs. Table 5 shows the computational complexity of our model and DBCFace [27] with different size and batch-size inputs. Our model is with a little higher computational complexity than DBCFace, but our model gets much better performance when using the same backbone for feature extraction, which means our method balances performance and efficiency.

6. Conclusion

In this paper, aiming at the difficulties of face detection in scale variation, shape difference and image context complexity, we propose an improved anchor-free framework. In particular, an online scale adaption strategy is introduced to guide each detector to learn a best detection range of face scale. In addition, we propose a context-aware detection module to explicitly transform original features into the same feature space. Furthermore, shape-sensitive module is designed to deal with faces with singular aspect ratio. On four common challenging benchmarks, WIDER FACE,FDDB, AFW and PASCAL face datasets, extensive experiments demonstrate that our method achieves the state-of-the-art detection performance. Our method shows that anchor-based is no longer the only choice for high performance face detection, and anchor-free method can achieve the same or even higher performance on the task of face detection with large scale variation. Our proposed model could adapt to various face scales, the network could detect tiny face in large scale while capture big face in small scale. This technology could provide inspiration about the design of scale-robust face detection models. In intelligent monitoring, automatic driving and other tasks, it is necessary to detect all people and not miss any target, our method could be applied to these scenarios.

In the future, we will explore the transfer learning, domain adaptation and neural network architecture search techniques to improve the generalization and address the challenging problems in face detection like large scale variation, occlusion and very tiny face detection.

CRediT authorship contribution statement

Cunying Ye: Helps conduct many experiments and writes a part of this paper. **Xin Li:** Contributes the main idea of this work, and is the main hand in paper writing. **Shenqi Lai:** Gives many suggestions on paper writing and organization. **Yaxiong Wang:** Advices the paper writing and helps review the paper. **Xueming Qian:** Charge of supervising all steps, including the main idea, and paper writing and the experiments organization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported in part by the Science and Technology Program of Xi'an, China under Grant 21RGZN0017.

References

[1] J. Sun, W. Yang, J.-H. Xue, Q. Liao, An equalized margin loss for face recognition, IEEE Trans. Multimed. 22 (11) (2020) 2833–2843, http://dx.doi.org/10.1109/TMM.2020.2966863.

[2] F. Wang, J. Cheng, W. Liu, H. Liu, Additive Margin softmax for face verification, IEEE Signal Process. Lett. 25 (7) (2018) 926–930, http://dx.doi.org/10.1109/LSP.2018.2822810.

[3] S. Zhao, W. Liu, S. Liu, J. Ge, X. Liang, A hybrid-supervision learning algorithm for real-time un-completed face recognition, Comput. Electr. Eng. 101 (2022) 108090, http://dx.doi.org/10.1016/j.compeleceng.2022.108090, URL https://www.sciencedirect.com/science/article/pii/S0045790622003457.

[4] J. Lv, X. Shao, J. Xing, C. Cheng, X. Zhou, A deep regression architecture with two-stage re-initialization for high performance facial landmark detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 3691–3700, http://dx.doi.org/10.1109/CVPR.2017.393.

[5] P. Gao, K. Lu, J. Xue, L. Shao, J. Lyu, A coarse-to-fine facial landmark detection method based on self-attention mechanism, IEEE Trans. Multimed. 23 (2021) 926–938, http://dx.doi.org/10.1109/TMM.2020.2991507.

[6] J. Wan, Z. Lai, J. Liu, J. Zhou, C. Gao, Robust face alignment by multi-order high-precision hourglass network, IEEE Trans. Image Process. 30 (2021) 121–133, http://dx.doi.org/10.1109/TIP.2020.3032029.

[7] C. Jing, Z. Dong, M. Pei, Y. Jia, Heterogeneous hashing network for face retrieval across image and video domains, IEEE Trans. Multimed. 21 (3) (2019) 782–794, http://dx.doi.org/10.1109/TMM.2018.2866222.

[8] Y.R. Choi, R.M. Kil, Face video retrieval based on the deep CNN with RBF loss, IEEE Trans. Image Process. 30 (2021) 1015–1029, http://dx.doi.org/10.1109/TIP.2020.3040847.

[9] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2879–2886.

[10] J. Yan, X. Zhang, Z. Lei, S.Z. Li, Face detection by structural models, Image Vis. Comput. 32 (10) (2014) 790–799.

[11] V. Jain, E. Learned-Miller, FDDB: A Benchmark for Face Detection in Unconstrained Settings, Technical Report, (UM-CS-2010-009) University of Massachusetts, Amherst, 2010.

[12] S. Yang, P. Luo, C.-C. Loy, X. Tang, Wider face: A face detection benchmark, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5525–5533.

[13] M. Najibi, P. Samangouei, R. Chellappa, L.S. Davis, SSH: Single stage headless face detector, in: 2017 IEEE International Conference on Computer Vision, ICCV, 2017, pp. 4885–4894, http://dx.doi.org/10.1109/ICCV.2017.522, URL https://doi.ieeecomputersociety.org/10.1109/ICCV.2017.522.

[14] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, S.Z. Li, S3fd: Single shot scale-invariant face detector, in: The IEEE International Conference on Computer Vision, ICCV, 2017, pp. 192–201.

[15] J. Wang, Y. Yuan, G. Yu, Face attention network: An effective face detector for the occluded faces, 2017, ArXiv, arXiv:1711.07246.

[16] X. Tang, D.K. Du, Z. He, J. Liu, Pyramidbox: A context-assisted single shot face detector, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 797–813.

[17] C. Chi, S. Zhang, J. Xing, Z. Lei, S.Z. Li, X. Zou, Selective refinement network for high performance face detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 8231–8238.

[18] T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 936–944, http://dx.doi.org/10.1109/CVPR.2017.106, URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.106.

[19] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, S.Z. Li, Faceboxes: A CPU real-time face detector with high accuracy, in: 2017 IEEE International Joint Conference on Biometrics, IJCB, IEEE, 2017, pp. 1–9.

[20] J. Liang, J. Wang, Y. Quan, T. Chen, J. Liu, H. Ling, Y. Xu, Recurrent exposure generation for low-light face detection, IEEE Trans. Multimed. (2021) 1, http://dx.doi.org/10.1109/TMM.2021.3068840.

[21] L. Huang, Y. Yang, Y. Deng, Y. Yu, Densebox: Unifying landmark localization with end to end object detection, 2015, arXiv preprint arXiv:1509.04874.

[22] J. Yu, Y. Jiang, Z. Wang, Z. Cao, T. Huang, Unitbox: An advanced object detection network, in: Proceedings of the 24th ACM International Conference on Multimedia, ACMPress, 2016, pp. 516–520.

[23] Y. Xue, Y. Li, S. Liu, X. Zhang, X. Qian, Crowd scene analysis encounters high density and scale variation, IEEE Trans. Image Process. 30 (2021) 2745–2757, http://dx.doi.org/10.1109/TIP.2021.3049963.

[24] H. Law, J. Deng, Cornernet: Detecting objects as paired keypoints, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 734–750.

[25] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, Q. Tian, Centernet: Keypoint triplets for object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6569–6578.

[26] X. Zhou, J. Zhuo, P. Krahenbuhl, Bottom-up object detection by grouping extreme and center points, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 850–859.

[27] X. Li, S. Lai, X. Qian, DBCFace: Towards pure convolutional neural network face detection, IEEE Trans. Circu. Syst. Video Technol. 32 (4) (2022) 1792–1804, http://dx.doi.org/10.1109/TCSVT.2021.3082635.

[28] W. Zheng, M. Yue, S. Zhao, S. Liu, Attention-based spatial-temporal multi-scale network for face anti-spoofing, IEEE Trans. Biom. Behav. Identity Sci. 3 (3) (2021) 296–307, http://dx.doi.org/10.1109/TBIOM.2021.3066983.

[29] J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou, Retinaface: Single-shot multi-level face localisation in the wild, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5203–5212.

[30] R. Girshick, Fast R-CNN, in: 2015 IEEE International Conference on Computer Vision, ICCV, 2015, pp. 1440–1448, http://dx.doi.org/10.1109/ICCV.2015.169.

[31] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99.

[32] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.

[33] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: European Conference on Computer Vision, Springer, 2016, pp. 21–37.

[34] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7263–7271.

[35] L. Ding, Y. Wang, R. Laganière, D. Huang, X. Luo, H. Zhang, A robust and fast multispectral pedestrian detection deep network, Knowl.-Based Syst. 227 (2021) 106990, http://dx.doi.org/10.1016/j.knosys.2021.106990, URL https://www.sciencedirect.com/science/article/pii/S0950705121002537.

[36] C. Wang, Z. Luo, S. Lian, S. Li, Anchor free network for multi-scale face detection, in: 2018 24th International Conference on Pattern Recognition, ICPR, IEEE, 2018, pp. 1554–1559.

[37] C. Wang, Z. Luo, Z. Zhong, S. Li, SAFD: single shot anchor free face detector, Multimedia Tools Appl. 80 (9) (2021) 13761–13785.

[38] X. Feng, L. Duan, J. Chen, An automated method with anchor-free detection and U-shaped segmentation for nuclei instance segmentation, in: Proceedings of the 2nd ACM International Conference on Multimedia in Asia, 2021, pp. 1–6.

[39] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, pp. 483–499.

[40] T. Ma, W. Tian, P. Kuang, Y. Xie, An anchor-free object detector with novel corner matching method, Knowl.-Based Syst. 224 (2021) 107083, http://dx.doi.org/10.1016/j.knosys.2021.107083, URL https://www.sciencedirect.com/science/article/pii/S0950705121003464.

[41] Y. Yang, X. Tang, Y.-M. Cheung, X. Zhang, F. Liu, J. Ma, L. Jiao, AR2Det: An accurate and real-time rotational one-stage ship detector in remote sensing images, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–14, http://dx.doi.org/10.1109/TGRS.2021.3092433.

[42] W. Ma, T. Zhou, J. Qin, Q. Zhou, Z. Cai, Joint-attention feature fusion network and dual-adaptive NMS for object detection, Knowl.-Based Syst. 241 (2022) 108213, http://dx.doi.org/10.1016/j.knosys.2022.108213, URL https://www.sciencedirect.com/science/article/pii/S0950705122000582.

[43] X. Wang, S. Lai, Z. Chai, X. Zhang, X. Qian, SPGNet: Serial and parallel group network, IEEE Trans. Multimed. 24 (2022) 2804–2814, http://dx.doi.org/10.1109/TMM.2021.3088639.

[44] X. Wu, D. Hong, J. Chanussot, Y. Xu, R. Tao, Y. Wang, Fourier-based rotation-invariant feature boosting: An efficient framework for geospatial object detection, IEEE Geosci. Remote Sens. Lett. 17 (2) (2020) 302–306, http://dx.doi.org/10.1109/LGRS.2019.2919755.

[45] X. Wu, D. Hong, J. Tian, J. Chanussot, W. Li, R. Tao, ORSIm detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features, IEEE Trans. Geosci. Remote Sens. 57 (7) (2019) 5146–5158, http://dx.doi.org/10.1109/TGRS.2019.2897139.

[46] X. Zhou, D. Wang, P. Krähenbühl, Objects as points, 2019, arXiv preprint arXiv:1904.0785.

[47] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, B. Zhang, More diverse means better: Multimodal deep learning meets remote-sensing imagery classification, IEEE Trans. Geosci. Remote Sens. 59 (5) (2021) 4340–4354, http://dx.doi.org/10.1109/TGRS.2020.3016820.

[48] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.

[50] C. Zhu, R. Tao, K. Luu, M. Savvides, Seeing small faces from Robust anchor's perspective, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5127–5136.

[51] C. Zhu, Y. He, M. Savvides, Feature selective anchor-free module for single-shot object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 840–849.

[52] Z. Tian, C. Shen, H. Chen, T. He, Fcos: Fully convolutional one-stage object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9627–9636.

[53] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, J. Shi, FoveaBox: Beyound anchor-based object detection, IEEE Trans. Image Process. 29 (2020) 7389–7398, http://dx.doi.org/10.1109/TIP.2020.3002345.

[54] J. Zhang, X. Wu, S.C. Hoi, J. Zhu, Feature agglomeration networks for single stage face detection, Neurocomputing 380 (2020) 180–189.

[55] P. Hu, D. Ramanan, Finding tiny faces, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017.

[56] K. Zhang, Z. Zhang, H. Wang, Z. Li, Y. Qiao, W. Liu, Detecting faces using inside cascaded contextual cnn, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2017, pp. 3171–3179.

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

[58] D. Triantafyllidou, P. Nousi, A. Tefas, Fast deep convolutional face detection in the wild exploiting hard sample mining, Big Data Res. 11 (2018) 65–76.

[59] E. Ohn-Bar, M.M. Trivedi, To boost or not to boost? On the limits of boosted trees for object detection, in: 23rd International Conference on Pattern Recognition, ICPR 2016, CancÚN, Mexico, December 4-8, 2016, IEEE, 2016, pp. 3350–3355, http://dx.doi.org/10.1109/ICPR.2016.7900151.

[60] D. Triantafyllidou, A. Tefas, A fast deep convolutional neural network for face detection in big visual data, in: INNS Conference on Big Data, Springer, 2016, pp. 61–70.

[61] R. Ranjan, V.M. Patel, R. Chellappa, HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2018) 1.