Multimedia Tools and Applications (2023) 82:10535–10552
https://doi.org/10.1007/s11042-022-13667-5

Wide aspect ratio matching for robust face detection

Shi Luo[1,2] · **Xiongfei Li**[1,2] · **Xiaoli Zhang**[1,2]

Received: 10 March 2021 / Revised: 30 April 2022 / Accepted: 11 August 2022 /
Published online: 6 September 2022

Abstract
Recently, anchor-based methods have achieved great progress in face detection. They adopt standard anchor matching strategy to sample positive anchors according to predefined IoU threshold. However, the max IoUs of extreme aspect ratio faces are still lower than fixed positive threshold, leading to the sampling failure from these faces. To construct a more robust detection model, more positive anchors from extreme aspect ratio faces need to be sampled and participate in the training phase. The goal of the present research is to improve the detection performance by reasonably extending sampling range of face aspect ratio. In this paper, we firstly explore the factors that affect the max IoU of each face in theory. Then, anchor matching simulation is performed to evaluate the sampling range of face aspect ratio. Finally, we propose a Wide Aspect Ratio Matching (WARM) strategy to collect more representative positive anchors from ground-truth faces across a wide range of aspect ratios. Besides, we present a novel feature enhancement module, named Receptive Field Diversity (RFD) module, to provide diverse receptive field corresponding to different aspect ratios. Extensive experiments have been conducted on popular benchmarks to show the effectiveness of our method, which can help detectors better capture extreme aspect ratio faces. Our method achieves promising APs on WIDER FACE validation dataset (easy: 0.965, medium: 0.955, hard: 0.904) and impressive generalization capability on FDDB dataset.

Keywords Face detection · Anchor matching · Feature enhancement · Deep learning · Convolutional neural network

1 Introduction

Accurate face detection is a prerequisite of many face related applications, such as face alignment [17, 42], face recognition [7, 27, 41] and facial emotion recognition [1], etc. Moreover, face detection and its extension technology [35] are widely used in various real-life scenarios to prevent the spread of the COVID-19 virus. Limited by the size of dataset

✉ Xiaoli Zhang
zhangxiaoli@jlu.edu.cn

[1] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China

[2] College of Computer Science and Technology, Jilin University, Changchun 130012, China

and hardware conditions, traditional methods design hand-crafted features to detect faces. In the era of deep learning, convolutional neural networks (CNN) have achieved remarkable successes in a variety of computer vision tasks, ranging from image classification [3, 11] to object detection [10], which also inspire face detection. CNN-based methods utilize popular backbone network (e.g., VGG [34], ResNet, DenseNet [13], and MobileNet) and feature fusion strategy like FPN [9] to extract face features from huge dataset automatically. Although it is difficult to explain the reasons from human perspective, the detection performance of these CNN-based methods is further improved indeed. Therefore, traditional methods are more efficiency and interpretable, while deep learning methods achieve higher detection accuracy.

Among them, anchor-based methods play a dominant role in CNN-based face detectors. During the training phase, a series of anchors are preset on the images with different scales and single aspect ratio (AR). Later, standard anchor matching (SAM) strategy [32] assigns positive and negative labels on these anchors according to intersection-over-union (IoU) threshold. We assign anchors to positive labels when their IoU threshold are in (T_p, 1], and negative labels if their IoU in [0, T_n). The rest anchors are discarded. Finally, these labeled anchors are feed into training network and update its parameters. During the inference phase, they detect faces by classifying and regressing these anchors. Predict score from classification branch can tell us if current anchor contain a face while coordinate offsets of current anchor locates its position from regression branch. Obviously, sufficient training samples and complex feature extraction networks are essential for high performance anchor-based face detection.

However, sampling positve anchors from each face is not always successful. SAM strategy utilizes the identical sampling threshold for all faces. When anchor sampling threshold fixed, the AR sampling range for positive anchors is determined. The ground-truth faces whose AR out of this range, called extreme aspect ratio faces, will be neglect. Figure 1

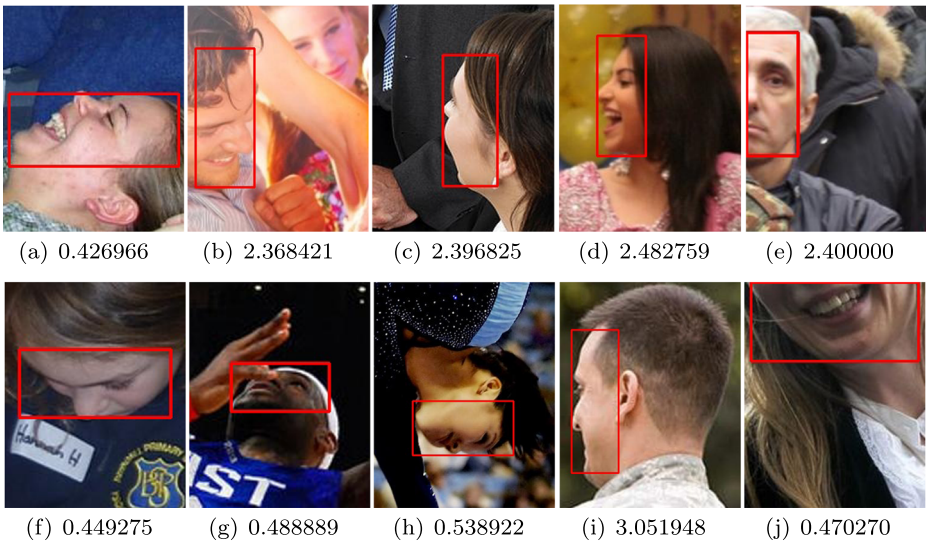


Fig. 1 Extreme aspect ratio faces on Wider Face training set. The aspect ratio (AR) of each ground truth face is noted blow. The first row faces fail to be sampled with anchor AR of 1.0 while the second row faces with anchor AR of 1.25. The first four columns show extreme pose faces while the last column are partial faces

shows some extreme AR faces on Wider Face [45] training set. We can see that extreme AR faces consist of extreme pose faces and partial faces. Extreme pose faces come from head rotation in pitch, yaw, and roll axis as seen in the first four columns of Fig. 1. And partial faces are derived from occlusion as shown in the last column of Fig. 1. To construct a more robust detection model, there is a rising demand for sampling more positive anchors from extreme aspect ratio faces. So, fixed sampling threshold for all faces is no longer suitable.

Furthermore, diversified aspect raitos conflict with single receptive field of network. Current feature enhancement modules usually adopt symmetric filters to enlarge receptive field. In fact, anchors with diverse receptive field make it easier to be classified and regressed in a convolutional manner.

In this paper, we firstly investigate what makes the max IoU of each face different and theoretically prove that AR of faces is the key impact. Then, anchor matching simulation is performed to evaluate the sampling range of face AR. Obviously, the sampling failure from extreme AR faces is just because of their ARs out of sampling range. In fact, the max IoUs of these extreme AR faces are still lower than fixed sampling threshold in SAM strategy. Therefore, we propose a Wide Aspect Ratio Matching (WARM) strategy to collect more representative positive anchors from ground-truth faces with a wide range of ARs. Specifically, extreme AR faces have their own sampling threshold according to AR themselves. Besides, we design a novel feature enhancement module, named Receptive Field Diversity (RFD) module, to provide large receptive field and diverse aspect ratio simultaneously. Both symmetric and asymmetric convolution kernels are used. Finally, we conduct comprehensive experiments on popular benchmarks, including WIDER FACE and FDDB [14] datasets, and achieve promising detection performance, especially for extreme AR faces.

For clarity, the main contributions of this paper can be summarized as:

(1) We theoretically prove that aspect ratio of faces is the key factors affecting the max IoU of each face and perform anchor matching simulation to evaluate the sampling range of face aspect ratio.
(2) We propose a WARM strategy to collect more representative positive anchors from ground-truth faces with a wide range of aspect ratio.
(3) A novel feature enhancement module, named RFD module, is designed to provide more diverse aspect ratio and large receptive field simultaneously.
(4) Our method can help detectors better capture extreme aspect ratio faces and achieve promising detection performance on challenging face detection benchmarks, including WIDER FACE and FDDB datasets.

The rest of the paper is organized as follows. Section 2 briefly reviews the related work in face detection. Section 3 presents our proposed method. Section 4 shows our experimental results. Section 5 concludes this paper.

2 Related work

Face detection is a fundamental step to various face related applications. Previous methods can be roughly divided into two categories as follows:
Traditional method: They detect faces based on hand-crafted features (e.g., HOG, LBP, HAAR, SIFT, SURF, and ORB [2]) in a sliding-window manner and optimize each component separately. The pioneering work of Viola-Jones [40] used Haar-like features and

AdaBoost to train a cascade of face detectors. LBP [16] introduced local texture features for face detection. NPDFace [22] presented normalized pixel difference feature to address challenges in unconstrained face detection, such as arbitrary pose and heavy occlusion. Kumar et al. [18] combined YCbCr, HSV and L × a × b color model to detect faces from still images under occlusion and non-uniform illumination.

CNN-based method: Different from traditional methods, CNN-based methods can automatically extract discriminative features from challenging face datasets. CascadeCNN [19] developed a cascade framework via deep CNNs to detect face coarse to fine. Faceness [44] trained a series of CNNs for facial attribute recognition to detect partially occluded faces. MTCNN [51] jointly solved face detection and alignment using several multi-task CNNs. HR [12] built multi-level image pyramids to boost the performance on extreme scale variations. UnitBox [47] presented an IoU loss to directly regress the bounding box. FANet [50] created a new hierarchical effective feature pyramid with rich semantics at all scales. BFBox [25] designed a FPN-attention module to joint search the face-appropriate space of backbone and FPN.

Recently, anchor-based methods have attracted more attention duo to their detection accuracy as well as inference efficiency. Effective anchor design and anchor matching strategy are necessary to generate more representative training samples. Besides, feature enhancement can further improve the ability of discriminative face features. Therefore, we focus on reviewing the prior works from this three perspectives below:

Anchor design In face detection, the choice of anchors and their placement on the image is very important. For example, using extra strided anchors are shown to be beneficial. ZCC [55] introduced a novel anchor design to guarantee high overlaps between faces and anchor boxes. PyramidBox [38] formulated a data-anchor-sampling strategy to increase the proportion of small faces in the training data. FaceBoxes [53] presented a new anchor densification strategy to improve the recall rate of small faces. FA-RPN [30] proposed an efficient anchor placement strategy to reduce the number of anchors to detect faces. Group sampling [28] emphasized the importance of balanced training samples, including both positive and negative ones, at different scales. In this paper, we continue to follow these guidances to form a high recall ratio anchor design, which adopts a wider range of anchor size and a shorter anchor stride.

Anchor matching SAM strategy utilized fixed sampling threshold to assign positive anchors. S^3FD [54] proposed scale compensation anchor matching strategy which helps the outer faces match more anchors. SRN [6] introduced a selective two-step classification to ignore training easy sample anchors in the second stage. DSFD [21] offered an improved anchor matching method to provide better initialization for the regressor. HAMBox [26] helped outer faces compensate high-quality anchors, which can obtain high IoU regression bounding boxes. Although anchor matching has been extensively studied, the failure of sampling positive anchors from extreme AR faces is still neglect. In this paper, we propose WARM strategy to sample more positive anchors from extreme AR faces.

Feature enhancement SSH [29] added large filters on each detection module to merge the context information. DSFD [21] introduced a feature enhance module to extend the single shot detector to dual shot detector. OS-LFFD [43] designed a novel ommateum block to maintain the corresponding ratio of receptive fields to face regions. RefineFace [48] constructed a RFE module to provide more diverse receptive fields for detecting extreme-pose faces. In this paper, we propose RFD module to adapt the diversity of face AR. Inspired

by ACNet [8], both symmetric and asymmetric convolution kernels are used into proposed RFD module.

3 Proposed method

This section introduces our proposed WARM strategy and RFD module. We firstly explore the factors of max IoU between anchors and each face. Next, anchor matching simulation is performed to evaluate sampling range of face aspect ratio. Then, we demonstrate WARM strategy in detail. Finally, we design RFD module to fit for face features with various aspect ratios.

3.1 Factors of face max IoU

For anchor-based face detection, one of the most important steps is to match ground-truth boxes with well-designed anchors and assign those anchors with positive and negative labels based on their IoUs. However, we discover that sampling positve anchors from each face is not always successful. When anchor design determined, SAM strategy always fails to sample positive anchors from faces with extreme AR. Therefore, we theoretically explore what affects the max IoU of each face and the reason of sampling failure.

Formally, the set of anchors is denoted as $A=\{a_i\}_{i=1}^{m}$, where i is the index of anchors and m is the number of anchors for all scales. $B_i^a = (x_i^a, y_i^a, w_i^a, h_i^a)$ is the bounding box of anchor a_i, where (x_i^a, y_i^a) is the upper-left coordinates and (w_i^a, h_i^a) represents the width and height of this anchor. Besides, r^a is the aspect ratio of all anchors. Similarly, the set of ground-truth faces is denoted as $G=\{g_j\}_{j=1}^{n}$, where j is the index of the ground-truth faces and n is the number of these faces. $B_j^g = (x_j^g, y_j^g, w_j^g, h_j^g)$ is the bounding box of ground-truth face g_j, where (x_j^g, y_j^g) is the upper-left coordinates and (w_j^g, h_j^g) represents the width and height of this face. In addition, r_j^g is the aspect ratio of face g_j. Given a ground-truth face g_j, the max IoU of this face can be computed as in (1).

$$\max IoU(g_j) = \max_{a_i \in A} \frac{B_j^g \cap B_i^a}{B_j^g \cup B_i^a} = \max_{a_i \in A} \frac{1}{\frac{Area(B_j^g)+Area(B_i^a)}{B_j^g \cap B_i^a} - 1}, \tag{1}$$

where $\cap$ and $\cup$ denote the intersection and union of two boxes respectively. Area(.) calculates the area of the bounding box. Furthermore, we recognize that there exists the max value of $B_j^g \cap B_i^a$ formulated as in (2).

$$\begin{aligned} \max(B_j^g \cap B_i^a) &= \min(w_j^g, w_i^a) * \min(h_j^g, h_i^a) \\ &= \min(w_j^g, w_i^a) * \min(w_j^g * r_j^g, w_i^a * r^a), \end{aligned} \tag{2}$$

When anchor design determined, the width w_j^g and AR r_j^g of ground-truth face g_j are the rest factors affecting face maxIoU(g_j). Moreover, anchor setting is well designed in a multi-scale manner. Hence, the AR of faces is the key factors to determine their max IoUs. In other words, different faces have their own max IoU overlap according to AR themselves. Unfortunately, SAM strategy utilizes the same IoU threshold for all faces. As a result, the failure of sampling positive anchors from extreme AR faces is just because the max IoUs of these faces are still lower than sampling IoU threshold.

3.2 Anchor matching simulation

According to the max IoU of faces as discussed in Section 3.1, Anchor Matching Simulation (AMS) is performed on Wider Face training set to evaluate the sampling range of face aspect ratio. We assume that enought random crop is executed. Thus, each face will have the chance to match anchors with its max IoU. The AMS can be described in the following steps:

- Step 1: Construct a high-recall anchor design.
- Step 2: Calculate max IoU of each face.
- Step 3: Judge if current face can match positive anchors.
- Step 4: Record the sampling range of face aspect ratio.

To be more specific, we firstly construct a high-recall anchor design. The anchor size ranges from 4 to 512 pixels while the anchor stride is $\sqrt{2}$. All anchors have the same AR. Next, the max IoU of each face can be calculated as shown in (1) and (2). Then, we compare face max IoU with positive sampling threshold T_p. If the max IoU of current face is greater than T_p, the positive anchors related to this face can be added into training samples. Finally, we update the sampling range of face AR.

The result of AMS is listed in Table 1. Here we take SAM strategy as an example. From the first three rows, we can see that sampling range centers around the AR of anchors. More experimental results show that anchor AR of 1.0 can achieve higher performance. Besides, we find that the sampling range enlarges gradually as positive sampling threshold reduces as listed in the last four rows of Table 1. For convenience, we define Aspect Ratio Sampling Domain (ARSD) to approximately describe the sampling range of face AR as follow:

Definition 1 (Aspect Ratio Sampling Domain). Given an anchor design A, where r^a is the aspect ratio of anchors. M denotes the anchor matching strategy. For a ground-truth face set G, the aspect ratio sampling domain D(r^a, η) is difined as

$$D(r^a, \eta) = \{x | r^a/\eta < x < r^a\eta\}, \tag{3}$$

where η is the radius of sampling domain. Thus, the left and right ARSD can be descirbed as follows:

$$D^-(r^a, \eta) = \{x | r^a/\eta < x < r^a\}, \tag{4}$$

$$D^+(r^a, \eta) = \{x | r^a \leq x < r^a\eta\}, \tag{5}$$

When positive sampling threshold T_p of SAM is set to 0.5, we notice that the ARSD is D(1.00,2.25) as listed in the third row of Table 1. However, statistics show that the AR

Table 1 The AMS is performed with different positive sampling threshold T_p and anchor aspect ratio R^a

T_p	R^a	Range	ARSD
0.50	1.50	0.666667 ~ 3.363636	D(1.50,2.25)
0.50	1.25	0.560000 ~ 2.809524	D(1.25,2.25)
0.50	1.00	0.449275 ~ 2.241379	D(1.00,2.25)
0.45	1.00	0.388889 ~ 2.586207	D(1.00,2.59)
0.40	1.00	0.333333 ~ 3.055556	D(1.00,3.06)
0.35	1.00	0.285714 ~ 3.666667	D(1.00,3.67)

The aspect ratio sampling range of matched faces is listed blow. We can approximately describe the sampling range as ARSD

of more than 99.96% faces on the Wider Face training set is in D(1.00,5.00). Therefore, a considerable part of extreme AR faces is neglected during anchor matching phase as seen in Fig. 1.

3.3 Wide aspect ratio matching

Current anchor matching strategy usually consists of two steps: SAM and anchor compensation. Each face firstly attemps to match all anchors with IoU higher than predefined threshold. However, some of these ground-truth faces may be unmatched in this step, especially for extreme AR faces. Then, unmatched faces will be compensated with the rest anchors that have highest IoU with them in current iteration. Obviously, compensated anchors may reduce the detection performance since these anchors have lower IoU with unmatched faces. Therefore, we believe that current anchor matching strategy is neither flexible nor sufficient to match the anchors in face detection.

To address this issue, we propose a Wide Aspect Ratio Matching strategy to collect more representative positive anchors from a wide range of face aspect ratios. The core idea is to construct variable positive threshold for extreme AR faces. We firstly determine the sampling domain of extreme AR faces according to the result of AMS as follows.

$$E(\eta_1, \eta_0) = D(r^a, \eta_1) \setminus D(r^a, \eta_0) \tag{6}$$

where $D(r^a, \eta_1)$ represents the total sampling domain in our WARM strategy and $D(r^a, \eta_0)$ is a subset of this sampling domain. According to the result of AMS, η_0 is set to 2.0. After that, the difference set $E(\eta_1, \eta_0)$ of these two ARSDs is the sampling domain of extreme AR faces. Furthermore, the left and right sampling domain of extreme AR faces can be noted below.

$$E^-(\eta_1, \eta_0) = D^-(r^a, \eta_1) \setminus D^-(r^a, \eta_0) \tag{7}$$

$$E^+(\eta_1, \eta_0) = D^+(r^a, \eta_1) \setminus D^+(r^a, \eta_0) \tag{8}$$

Then, we construct a variable positive threshold function in the sampling domain of extreme AR faces while follow the SAM strategy in the rest sampling domain. For simplicity, the linear function is applied. The positive threshold of our WARM can be formulated as follow:

$$T_p = \begin{cases} T_0 - \delta * \theta(r_j^g), & r_j^g \in E(\eta_1, \eta_0) \\ T_0, & otherwise \end{cases} \tag{9}$$

where T_0 is the initial value of positive threshold and δ represents the amplitude of positive threshold. Similar to SAM, T_0 is set to 0.5. Besides, $\theta(r_j^g)$ reflects the change rate of positive threshold associated with the AR r_j^g of each face. When the AR of a face is far away from current anchor's, positive sampling threshold related to this face should decrease gradually. Thus, a simple implementation of $\theta(x)$ is given below.

$$\theta(x) = \begin{cases} \dfrac{\max(x) - x}{\max(x) - \min(x)}, & x \in E^-(\eta_1, \eta_0) \\ \dfrac{x - \min(x)}{\max(x) - \min(x)}, & x \in E^+(\eta_1, \eta_0) \end{cases} \tag{10}$$

To visualize our proposed WARM strategy, we plot the scatter diagram of all training faces in Fig. 2. The coordinate of each blue dot represents the AR of a face and its maximum IoU. It should be noted that face max IoU is the IoU value of the best matching anchor with this face. The green line is the positive threshold boundary of SAM while the red lines are our proposed WARM's. Specifically, line AB and CD represent the variable positive

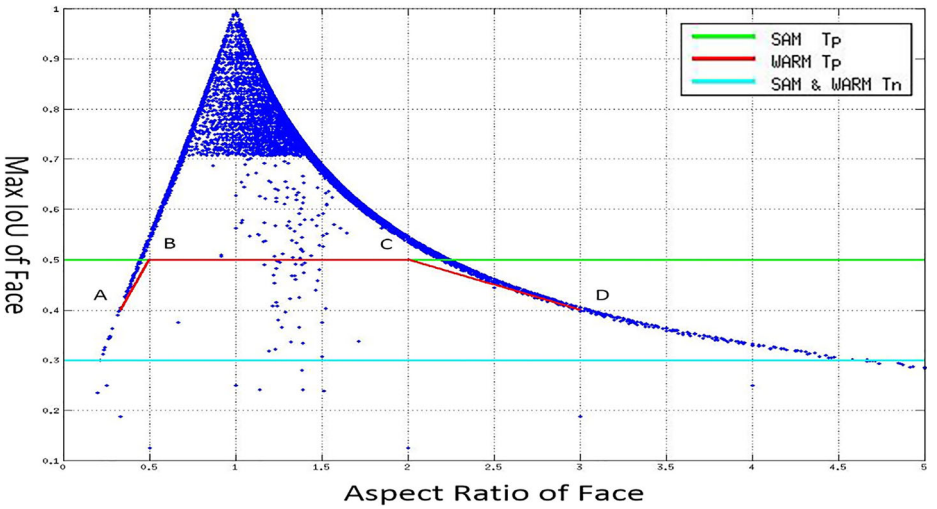


Fig. 2 The distribution of face aspect ratio and their max IoUs. The coordinate of blue dot represents the aspect ratio of each face and their max IoUs. The green line is the positive threshold boundary of SAM while the red lines are our proposed WARM's. Both of them adopt the cyan line as negative threshold boundary

threshold boundary for extreme AR faces. Instead of anchor compensation, variable positive threshold can match extreme AR faces with higher IoU anchors. Similarly, both of them adopt the cyan line as negative threshold boundary. When δ is set to 0, our proposed WARM degenerates to SAM strategy. Therefore, our proposed WARM method could be a general strategy for anchor-based single-stage face detection.

Although reducing positive sampling threshold in SAM can expand the sampling range of face AR as seen in Table 1, the average IoU of all positive anchors will decline, which is harmful for the quality of training samples. Besides, it is difficult to guarantee the high-quality positive anchors for unmatched faces can be compensated at each iteration. Different from SAM and anchor compensation, our proposed WARM strategy can sample high-quality positive anchors from extreme AR faces while basically maintain the whole quality of all positive samples at the same time.

3.4 Receptive field diversity

Current feature enhancement modules usually enlarge receptive field by mean of symmetric convolution kernels. The singleness of the receptive field is not fit for extreme AR faces. To address this issue, we propose a novel feature enhancement module, named Receptive Field Diversity (RFD) module, to provide diverse aspect ratio and large receptive field simultaneously. Both symmetric and asymmetric convolution kernels [8] are used to adapt the diversity of face AR.

Figure 3 illustrates the structure of RFD module, which is inspired by Inception [37] and ResNet blocks. The RFD module adopts a 4-path structure. In particular, we firstly utilize a 1x1 convolution layers to reduce the channel number to one quarter of the input feature maps. Then, the 3×3, 5×5, 3×1 and 1×3 convolution kernels are employed to provide diversity receptive field. Finally, these 4-path feature maps are concatenated togather. Besides, we apply a shortcut path, which maintain the original receptive field, to sum up with concatenated features above.

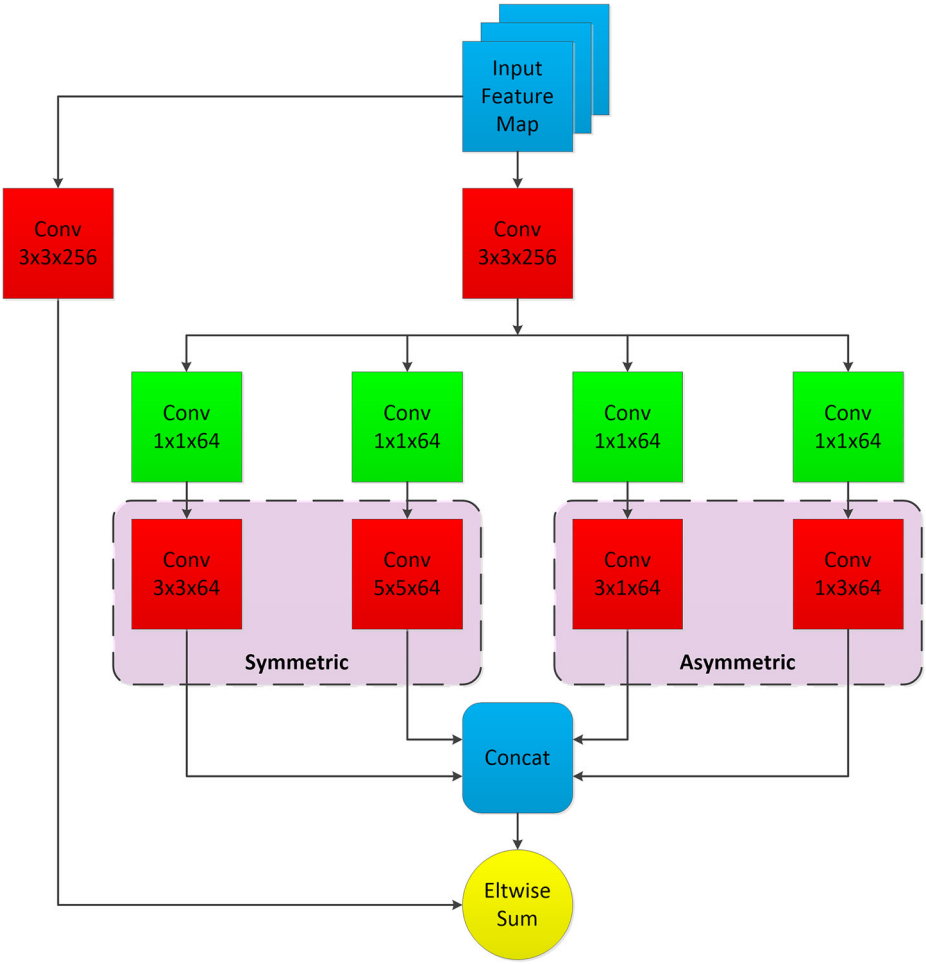


Fig. 3 The structure of Receptive Field Diversity module. Both symmetric and asymmetric convolution kernels are used to provide large receptive field and diverse aspect ratio simultaneously

4 Experiments

4.1 Experimental setup

In this section, we introduce the backbone, anchor design, data agumentation, loss function and other implementation details.

Backbone We adopt ResNet-50 with 5-level feature pyramid structure as the backbone network in our method. The feature maps are extracted from those four residual blocks, denoted as C2, C3, C4 and C5. P2, P3, P4, and P5 are the fused feature maps [9] corresponding to C2, C3, C4 and C5, while P6 is just down-sampled by two 3x3 convolution layers after C5.

Anchor design For anchor generation, we assign $\{4, 4\sqrt{2}, 8\}$ in P_2, $\{8\sqrt{2}, 16, 16\sqrt{2}\}$ in P_3, $\{32, 32\sqrt{2}, 64\}$ in P_4, $\{64\sqrt{2}, 128, 128\sqrt{2}\}$ in P_5, and $\{256, 256\sqrt{2}, 512\}$ in P_6. All anchors have aspect ratio of 1.0.

Data augmentation We randomly crop [38, 54] square patches from the original images and resize these patches into 640×640. Except for random crop, we also utilize random horizontal flip with probability of 0.5 and photo-metric color distortion [54] to augment training data.

Loss function We apply the multi-task loss as our objective function. Specifically, Focal loss [23] is used for the binary classification while Smooth-L1 loss is for the bounding box regression.

Optimization details We use stochastic gradient descent (SGD) with momentum 0.9 and weight decay 5×10^{-5} to fine-tune our detection models. The learning rate is set to 0.01 for the first 60 epochs, and decreases to 10^{-3} and 10^{-4} for the next two 30 epochs. Besides, OHEM [33] is applied to alleviate significant imbalance between the positive and negative training examples with a ratio of 1:3. During training phase, 256 detections per module are selected for each image. During inference phase, each module outputs 1000 detection results whose confidence scores are all higher than the threshold of 0.02. Finally, we perform NMS with a threshold of 0.3 on the outputs of all modules together. Our method is implemented in MXNet [5] and all the experiments are trained on 2 NVIDIA GeForce GTX 1080Ti GPUs in parallel.

4.2 Datasets

WIDER FACE dataset: It consists of 32,203 images with 393,703 labeled face boxes with a high degree of variability in scale, pose and occlusion. These images are split into training (40%), validation (10%), and testing (50%) sets by randomly sampling from 61 event classes. Faces in this dataset are classified into Easy, Medium, and Hard subsets according to their detection difficulty. We train all models on the training set of the WIDER FACE dataset while evaluate on its validation and test sets. Ablation studies are also performed on the validation set.

FDDB dataset: It contains 2845 images and 5171 annotated faces. Most of these faces have large scale, high resolutions or slightly occlusion sometimes. Different from WIDER FACE, faces in the FDDB dataset are labeled by bounding ellipses. In order to verify generalization ability of our method, we perform the evaluation on the FDDB dataset.

4.3 Ablation study

In this subsection, we conduct ablation studies to evaluate the effectiveness of our proposed WARM and RFE. For fair comparisons, we use the same settings as described in Section 4.1 for all the experiments.

The effect of wide aspect ratio matching strategy We discuss the effect of two hyper-parameters in our proposed WARM strategy. The performance under different η_1, δ (defined in Section 3.3), is shown in Table 2. Compared with SAM, our proposed WARM can collect

Table 2 Varying η_1, δ for WARM on WIDER FACE validation set

Method	η_1	δ	AP		
			Easy	Medium	Hard
SAM	2.25	0.00	0.959	0.950	0.898
	2.50	0.05	0.962	0.952	0.899
	2.50	0.10	0.961	0.952	0.901
WARM	2.50	0.15	0.958	0.949	0.897
	3.00	0.10	0.962	0.953	0.902
	4.00	0.10	0.960	0.951	0.898

more positive anchors from extreme AR faces, whose ARSD is in E(η_1,2.0). Besides, these extra collected positive anchors have higher IoU, ranging from 0.5-δ to 0.5, related to their matched extreme AR faces. After multiple ablative experiments, we find the optimal hyper-parameters. When η_1 and δ are set to 3.0 and 0.1, our proposed WARM can increase the detection performance of 0.3%(Easy), 0.3%(Medium), and 0.4%(Hard) separately.

The effect of receptive field diversity module To demonstrate the effectiveness of RFD module, we conduct the comparation experiment between SSH and RFD module as shown in Table 3. Previous feature enhancement modules utilize symmetric convolutional kernels (e.g., 3×3, 5×5, and 7×7) to enlarge the receptive field. Here, we take SSH detection module as an example. In order to adapt the diversity of face AR, both the symmetric and asymmetric convolution kernels are applied in our proposed RFD module as seen in Fig. 3. Note that, the 3×3, 5×5, 3×1 and 1×3 convolution kernels are used to provide diversity receptive field. From the Table 3, we can see that RFD module can further enhance the feature maps and improve the detection performance of 0.2%(Easy), 0.1%(Medium), and 0.2%(Hard) respectively.

Combining the WARM strategy and RFD module, our method achieves a promising detection performance as shown in the last row of Table 3.

4.4 Evaluation on benchmark

We evaluate our proposed method against state-of-the-art methods on two public face detection benchmarks.

Table 3 Effectiveness of RFD module on the AP performance

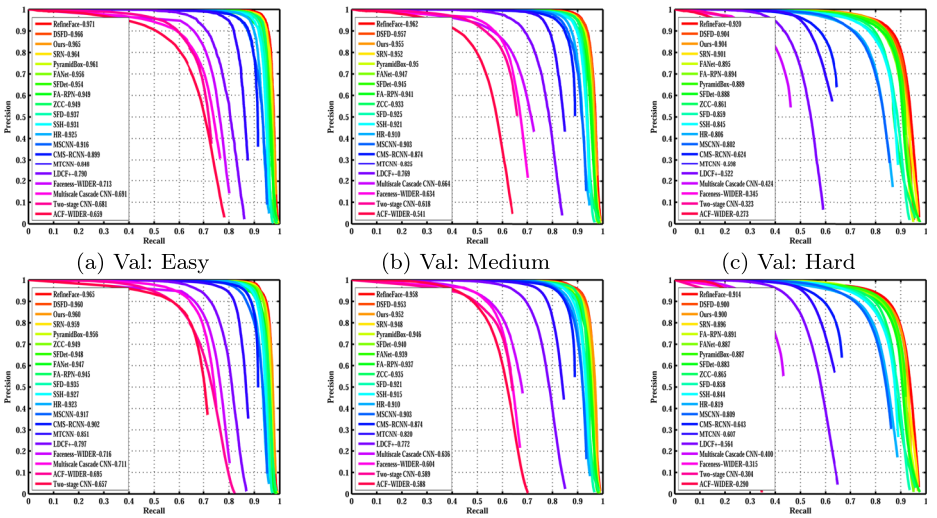
Component	AP		
	Easy	Medium	Hard
SSH	0.959	0.950	0.898
RFD	0.961	0.951	0.900
Ours(WARM + RFD)	0.965	0.955	0.904

4.4.1 Wider face dataset

Our model is trained on the training set and evaluate on its validation and testing set against the recently published state-of-the-art face detection methods including RefineFace [48], DSFD [21], SRN [6], PyramidBox [38], FANet [50], SFDet [49], FA-RPN [30], ZCC [55], S^3FD [54], SSH [29], HR [12], MSCNN [4], CMS-RCNN [56], MTCNN [51], LDCF+ [31], Faceness [44], Multiscale Cascade CNN [45], ACF [46] and Two-stage CNN [45]. The precision-recall curves and AP values on WIDER FACE validation and testing sets are shown in Fig. 4. Our method based on resnet-50 achieves the promising average precision in all level faces, i.e., 0.965 (Easy), 0.955 (Medium), 0.904 (Hard) for validation set, and 0.960 (Easy), 0.952 (Medium), 0.900 (Hard) for testing set. It can be seen that our method outperforms almost all other methods, demonstrating the effectiveness of our proposed method. Note that RefineFace and DSFD adopt resnet-152 as their backbone to achieve the detection performance as seen in Fig. 4. We believe that a deeper backbone like resnet-152 can further improve the detection performance of our method. More importantly, our study leads to a new insight to make detection models robust for extreme AR faces.

4.4.2 FDDB dataset

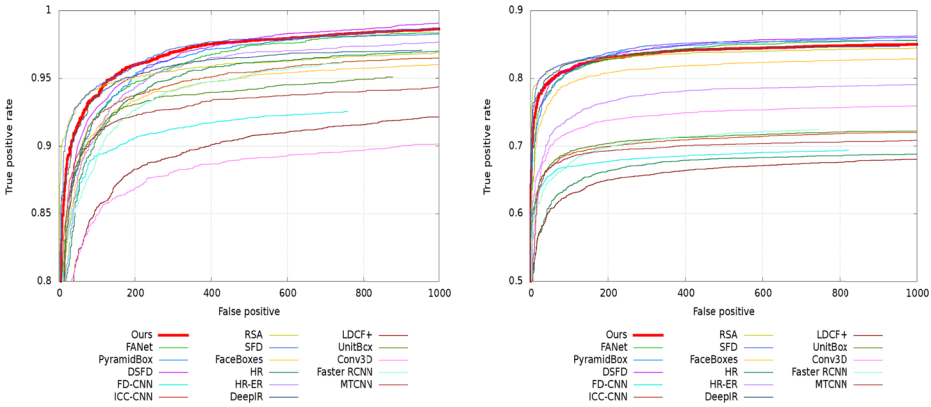
We directly use the same detection model above to perform the evaluation on FDDB dataset. Specifically, the shortest side of the input images is set to 400 pixels while the larger side is less than 800 pixels. We compare our method against the recently published state-of-the-art methods including FANet [50], PyramidBox [38], DSFD [21], FD-CNN [39], ICC-CNN [52], RSA [24], S^3FD [54], FaceBoxes [53], HR [12], HR-ER [12], DeepIR [36], LDCF+ [31], UnitBox [47], Conv3D [20], Faster RCNN [15] and MTCNN [51] on FDDB dataset. For a more fair comparison, the predicted bounding boxes are converted to bounding ellipses. Figure 5 shows the discrete ROC curves and continuous ROC curves of these



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

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

Fig. 4 Precision-recall curves on WIDER FACE validation and test sets



(a) Discontinuous ROC curves (b) Continuous ROC curves

Fig. 5 Evaluation on the FDDB dataset

methods on the FDDB dataset respectively. As can be seen, our proposed method consistently achieves a relatively higher performance in terms of both the discrete ROC curves and continuous ROC curves. These results demonstrate the effectiveness and impressive generalization capability of our proposed method.

4.5 Qualitative results

Figure 6 shows some detection results of our proposed method on the WIDER FACE validation dataset. Our method is able to detect faces with different AR, especially for extreme AR faces. The detection results for extreme pose faces are shown in the first row of Fig. 6. Besides, our method can also detect partial faces caused by occlusion as seen in the last row of Fig. 6. Surprisingly, our detection model can capture some extra extrmeme AR faces which are missing labels.

Figure 7 shows some detection results generated by our detection model on the FDDB dataset. Benefit from excellent performance of our method in detecting extreme AR faces,

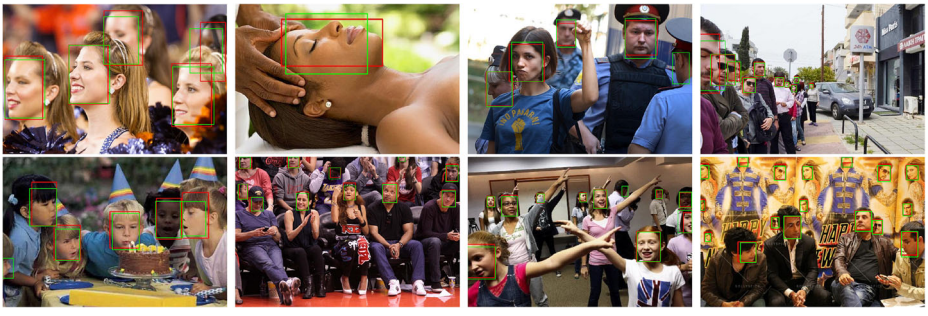


Fig. 6 Qualitative results on the WIDER FACE validation set. Red bounding boxes are the faces that annotated on the WIDER FACE validation dataset. Green bounding boxes represent the detection results. Best viewed in color. Please zoom in to see some small detections

Fig. 7 Qualitative results on the FDDB dataset. Red bounding ellipses are the faces that FDDB labeled; Green bounding boxes are the detection results. Best viewed in color. Please zoom in to see some small detections

we can find some more faces from human perspective but lack of labels on the FDDB dataset. Similarly, the detection results, which contain atypical pose and heavy occlusion faces, are presented in the first and last row of Fig. 7 separately.

Figure 8 is a qualitative result in world's largest selfie. Our method successfully find 911 faces out of the reported 1000 faces in the above image.

5 Conclusions and future work

In this paper, we examined the failure of sampling positive anchors from extreme AR faces and identified that the max IoUs of these faces are still lower than fixed sampling threshold

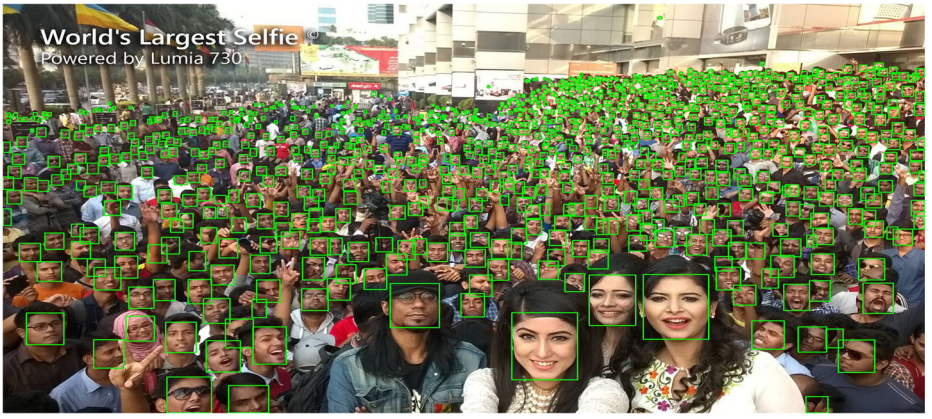


Fig. 8 The qualitative result in world's largest selfie. Our method successfully find 911 faces out of the reported 1000 faces in the above image. Best viewed in color. Please zoom in to see some small detections

in SAM strategy. Motivated by this observation, both Wide Aspect Ratio Matching strategy and Receptive Field Diversity module were deployed in our method for the sake of better detecting faces with different aspect ratio. These two strategies make our model effective and robust to detect faces with diversified AR in unconstrained settings, especially for extreme AR faces. Extensive experiments demonstrate that our method outperforms most of the recently published face detectors and achieves promising performance on challenging face detection benchmarks like WIDER FACE and FDDB datasets.

In the future, we attempt to construct nonlinear positive sample threshold boundary for extreme AR faces, to further improve the detection performance. Note that our proposed WARM strategy can flexibly adjust the sampling domain of aspect ratio, which can help detection models fit for the specific demand in real applications. Moreover, the design idea of WARM strategy proposed in this paper can also be transferred into other anchor-based detection task, which is also a direction of our future work.

Acknowledgements The work was supported in part by the National Natural Science Foundation of China under Grant 61801190, in part by the National Key Research and Development Project of China under Grant 2019YFC0409105, in part by the Nature Science Foundation of Jilin Province under Grant 20180101055JC, in part by the Industrial Technology Research and Development Funds of Jilin Province under Grant 2019C054-3, in part by the "Thirteenth Five-Year Pla" Scientific Research Planning Project of Education Department of Jilin Province (JKH20200678KJ,JJKH20200997KJ), and in part by the Fundamental Research Funds for the Central Universities, JLU.

Declarations

Conflict of Interests The authors declare that they have no conflict of interest.

References

1. Arora M, Kumar M (2021) AutoFER: PCA and PSO based automatic facial emotion recognition. Multimed Tools Appl 80(2):3039–3049
2. Bansal M, Kumar M, Kumar M (2021) 2D object recognition: a comparative analysis of SIFT, SURF and ORB feature descriptors. Multimed Tools Appl 80(12):18839–18857
3. Bansal M, Kumar M, Sachdeva M, Mittal A (2021) Transfer learning for image classification using VGG19: Caltech-101 image data set. J Ambient Intell Human Comput, https://doi.org/10.1007/s12652-021-03488-z
4. Cai Z, Fan Q, Feris RS, Vasconcelos N (2016) A unified multi-scale deep convolutional neural network for fast object detection. In: Proceedings of European conference on computer vision, pp 354–370
5. Chen T, Li M, Li Y, Lin M, Wang N, Wang M, Xiao T, Xu B, Zhang C, Zhang Z (2015) Mxnet: a flexible and efficient machine learning library for heterogeneous distributed systems. arXiv:1512.01274
6. Chi C, Zhang S, Xing J, Lei Z, Li SZ, Zou X (2018) Selective refinement network for high performance face detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), pp 8231–8238
7. Deng J, Guo J, Xue N, Zafeiriou S (2019) Arcface: additive angular margin loss for deep face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4685–4694
8. Ding X, Guo Y, Ding G, Han J (2019) ACNet: strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks. In: Proceedings of the IEEE international conference on computer vision, pp 1911–1920
9. Dollár P, Appel R, Belongie S, Perona P (2014) Fast feature pyramids for object detection. IEEE Trans Pattern Anal Mach Intell 36(8):1532–1545
10. Girshick R (2015) Fast R-CNN. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448
11. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference in computer vision and pattern recognition, pp 770–778

12. Hu P, Ramanan D (2017) Finding tiny faces. In: Proceedings IEEE conference of computer vision and pattern recognition, pp 951–959
13. Huang G, Liu Z, van der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference in computer vision and pattern recognition, pp 4700–4708
14. Jain V, Learned-Miller E (2010) FDDB: a Benchmark for face detection in unconstrained settings. University of Massachusetts. Amherst Tech Rep UM-CS-2010-009 2(7):8
15. Jiang H, Learned-Miller E (2017) Face detection with the faster r-cnn. In: Proceedings of IEEE international conference on automatic face & gesture recognition, pp 650–657
16. Jin H, Liu Q, Lu H, Tong X (2015) Face detection using improved LBP under Bayesian framework. In: Proceedings of International Conference on Images and Graphics, pp 306–309
17. Jourabloo A, Ye M, Liu X, Ren L (2017) Pose-invariant face alignment with a single CNN. In: Proceedings of the IEEE international conference on computer vision, pp 3219–3228
18. Kumar A, Kumar M, Kaur A (2021) Face detection in still images under occlusion and non-uniform illumination. Multimed Tools Appl 80(10):14565–14590
19. Li H, Lin Z, Shen X, Brandt J, Hua G (2015) A convolutional neural network cascade for face detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5325–5334
20. Li Y, Sun B, Wu T, Wang Y (2016) Face detection with end-to-end integration of a ConvNet and a 3D model. In: Proceedings of European conference on computer vision, pp 420–436
21. Li J, Wang Y, Wang C, Tai Y, Qiang J, Yang J, Wang C, Li J, Huang F (2019) Dsfd: dual shot face detector. In: Proceedings IEEE conference of computer vision and pattern recognition, pp 5055–5064
22. Liao S, Jain AK, Li SZ (2016) A fast and accurate unconstrained face detector. IEEE Trans Pattern Anal Mach Intell 38(2):211–223
23. Lin T, Goyal P, Girshick R, He K, Dollar P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988
24. Liu Y, Li H, Yan J, Wei F, Wang X, Tang X (2017) Recurrent scale approximation for object detection in CNN. In: Proceedings of the IEEE international conference on computer vision, pp 571–579
25. Liu Y, Tang X (2020) BFBOx: searching face-appropriate backbone and feature pyramid network for robust face detector. In: Proceedings IEEE conference of computer vision and pattern recognition, pp 13568–13577
26. Liu Y, Tang X, Han J, Liu J, Rui D, Wu X (2020) HAMBOx: delving into mining high-quality anchors on face detection. In: Proceedings IEEE conference of computer vision and pattern recognition, pp 13043–13051
27. Liu W, Wen Y, Yu Z, Li M, Raj B, Song L (2017) Sphereface: deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6738–6746
28. Ming X, Wei FY, Zhang T, Chen D, Wen F (2019) Group sampling for scale invariant face detection. In: Proceedings IEEE conference of computer vision and pattern recognition, pp 3441–3451
29. Najibi M, Samangouei P, Chellappa R, Davis LS (2017) Ssh: single stage headless face detector. In: Proceedings of IEEE international conference on computer vision, pp 4875–4884
30. Najibi M, Singh B, Davis LS (2019) FA-RPN: floating region proposals for face detection. In: Proceedings IEEE conference of computer vision and pattern recognition, pp 7715–7724
31. Ohn-Bar E, Trivedi MM (2016) To boost or not to boost? On the limits of boosted trees for object detection. In: Proceedings of international conference on pattern recognition, pp 3350–3355
32. Ren S, He K, Girshick R, Sun J (2017) Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149
33. Shrivastava A, Gupta A, Girshick R (2016) Training region-based object detectors with online hard example mining. In: Proceedings IEEE conference computer vision and pattern recognition, pp 761–769
34. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
35. Singh S, Ahuja U, Kumar M, Kumar K, Sachdeva M (2021) Face mask detection using YOLOv3 and faster r-CNN models: COVID-19 environment. Multimed Tools Appl 80(13):19753–19768
36. Sun X, Wu P, Hoi SCH (2018) Face detection using deep learning: an improved faster RCNN approach. Neurocomputing 299:42–50
37. Szegedy C, Liu W, Jia Y et al (2015) Going deeper with convolutions. In: Proceedings of the IEEE international conference on computer vision, pp 1–9
38. Tang X, Du DK, He Z, Liu J (2018) Pyramidbox: a context-assisted single shot face detector. In: Proceedings of European conference on computer vision, pp 797–813
39. Triantafyllidou D, Nousi P, Tefas A (2018) Fast deep convolutional face detection in the wild exploiting hard sample mining. Big Data Res 11:65–76
40. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vis 57(2):137–154

41. Wang H, Wang Y, Zhou Z, Ji X, Gong D, Zhou J, Li Z, Liu W (2018) Cosface: large margin cosine loss for deep face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5265–5274
42. Wu W, Qian C, Yang S, Wang Q (2018) Look at boundary: a boundary-aware face alignment algorithm. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2129–2138
43. Xu DZ, Wu LF, He YH, Zhao Q, Jian M, Yan JC (2020) OS-LFFD: a light and fast face detector with ommateum structure. Multimed Tools Appl. https://doi.org/10.1007/s11042-020-09143-7
44. Yang S, Luo P, Loy CC, Tang X (2015) From facial parts responses to face detection: a deep learning approach. In: Proceedings of the IEEE international conference on computer vision, pp 3676–3684
45. Yang S, Luo P, Loy CC, Tang X (2016) Wider face: a face detection benchmark. In: Proceedings of the IEEE conference in computer vision and pattern recognition, pp 5525–5533
46. Yang B, Yan J, Lei Z, Li SZ (2014) Aggregate channel features for multi-view face detection. In: Proceedings of IEEE international joint conference on biometrics, pp 1–8
47. Yu J, Jiang Y, Wang Z, Cao Z, Huang T (2016) UnitBox: an advanced object detection network. In: Proceedings of international conference on multimedia, pp 516–520
48. Zhang S, Chi C, Lei Z, Li SZ (2020) RefineFace: refinement neural network for high performance face detection. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2020.2997456
49. Zhang S, Wen L, Shi H, Lei Z, Lyu SW, Li SZ (2019) Single-shot scale-aware network for real-time face detection. Int J Comput Vis 127(6):537–559
50. Zhang J, Wu X, Zhu J, Hoi SCH (2020) Feature agglomeration networks for single stage face detection. Neurocomputing 380:180–189
51. Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23(10):1499–1503
52. Zhang K, Zhang Z, Wang H, Li Z, Qiao Y, Liu W (2017) Detecting faces using inside cascaded contextual CNN. In: Proceedings of the IEEE international conference on computer vision, pp 3190–3198
53. Zhang S, Zhu X, Lei Z, Shi H, Wang X, Li SZ (2017) Faceboxes: a CPU real-time face detector with high accuracy . In: Proceedings of IEEE international joint conference on biometrics, pp 1–9
54. Zhang S, Zhu X, Lei Z, Shi H, Wang X, Li SZ (2017) S3FD: Single shot scale-invariant face detector. In: Proceedings of the IEEE International Conference on Computer Vision, pp 192–201
55. Zhu C, Tao R, Luu K, Savvides M (2018) Seeing small faces from robust anchor's perspective. In: Proceedings IEEE conference of computer vision and pattern recognition, pp 5127–5136
56. Zhu C, Zheng Y, Luu K, Savvides M (2017) CMS-RCNN: contextual multi-scale region-based CNN for unconstrained face detection. In: Deep learning for biometrics. Springer, Cham, pp 57–79

Shi Luo received his MSc degree in computer science and technology from Jilin University in 2016. He is currently a PhD candidate in college of computer science and technology, Jilin University. His research interests include computer vision, pattern recognition, especially for object detection, face detection.

Xiongfei Li received the BS degree in computer software in 1985 from Nanjing University, the MSc degree in computer software in 1988 from the Chinese academy of sciences, the PhD degree in communication and information system in 2002 from Jilin University. Since 1988, he has been a member of the faculty of the computer science and technology at Jilin University, Changchun, China. He is a professor of computer software and theory at Jilin University. He has authored more than 100 research papers. His research interests include data mining, intelligent network, image processing and analysis. Prof. Li is a member of the IEEE.

Xiaoli Zhang received the M.Sc. and PhD degree in computer science and technology from Jilin University, in 2012 and 2016. Since 2018, he has been a member of the faculty of the computer science and technology at Jilin University, Changchun, China. He is a professor of computer software and theory at Jilin University. He has published more than 20 papers in journals and conferences. His research interests include information fusion, algorithm evaluation, and data mining.