Wide Aspect Ratio Matching for Robust Face Detection

Shi Luo\textsuperscript{1,2} · Xiongfei Li\textsuperscript{1,2} · Xiaoli Zhang\textsuperscript{1,2}

Received: date / Accepted: date

Abstract Recently, anchor-based methods have achieved great progress in face detection. Once anchor design and anchor matching strategy determined, plenty of positive anchors will be sampled. However, faces with extreme aspect ratio always fail to be sampled according to standard anchor matching strategy. In fact, the max IoUs between anchors and extreme aspect ratio faces are still lower than fixed sampling threshold. 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. Besides, 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 ratio. Finally, 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 show that 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.

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

\textsuperscript{1} Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China

\textsuperscript{2} College of Computer Science and Technology, Jilin University, Changchun 130012, China
1 Introduction

Accurate face detection is a prerequisite of many face related applications, such as face alignment [1,2], face recognition [3,4,5] and facial expression recognition, etc. In recent years, convolutional neural networks (CNN) have achieved remarkable successes in face detection. Unlike traditional face detectors of hand-crafted features, CNN-based methods utilize popular backbone network(e.g. VGG [6], ResNet [7], DenseNet [8], MobileNet) and feature fusion strategy(e.g. FPN [9]) to extract robust face features automatically.

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 [10] 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.

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 shows some extreme AR faces on Wider Face [11] 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 (see (a)-(d) and (f)-(i)). And partial faces are derived from occlusion (see (e), (j)). For sampling more positive anchors related to extreme AR faces, fixed sampling threshold for all faces is no longer suitable.

Furthermore, diversified aspect raito of faces conflict with single receptive field of network. Current feature enhancement modules usually adopt square filter 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 failure of sampling extreme AR faces is just because of their AR 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
AR. 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 more diverse aspect ratio and large receptive field simultaneously. Both symmetric and asymmetric convolution kernels are used. Finally, we conduct comprehensive experiments on popular benchmarks, including WIDER FACE and FDDB [12] 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 factor 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 in a sliding-window manner and optimize each component separately. The pioneering work of Viola-Jones [13] uses Haar-like features and AdaBoost to train a cascade of face detectors. LBP [14] introduces local texture features for face detection. NPDFace [15] proposes normalized pixel difference feature to address challenges in unconstrained face detection, such as arbitrary pose and heavy occlusion. Sawat et al. [16] adopts hand-crafted as well as visual features together for improving robustness in face detection.

CNN-based method: Different from traditional methods, CNN-based methods can automatically extract discriminative features from challenging face datasets. CascadeCNN [17] develops a cascade framework via deep CNNs to detect face coarse to fine. Faceness [18] trains a series of CNNs for facial attribute recognition to detect partially occluded faces. MTCNN [19] jointly solves face detection and alignment using several multi-task CNNs. HR [20] builds multi-level image pyramids to boost the performance on extreme scale variations. UnitBox [21] presents an IoU loss to directly regress the bounding box. FANet [22] creates a new hierarchical effective feature pyramid with rich semantics at all scales. BFBox [23] designs an FPN-attention module to joint search the face-appropriate space of backbone and FPN.

Recently, anchor-based methods have attracted more attention due 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 these 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 [24] introduces a novel anchor design to guarantee high overlaps between faces and anchor boxes. PyramidBox [25] formulates a data-anchor-sampling strategy to increase the proportion of small faces in the training data. FaceBoxes [26] presents a new anchor densification strategy to improve the recall rate of small faces. FA-RPN [27] proposes an efficient anchor placement strategy to reduce the number of anchors to detect faces. Group sampling [28] emphasizes 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 utilizes fixed sampling threshold to assign positive anchors. S$^3$FD [29] proposes scale compensation anchor matching strategy which helps the outer faces match more anchors. SRN [30] introduces a selective two-step classification to ignore training easy sample anchors in the second stage. DSFD [31] offers an improved anchor matching method to provide better initialization for the regressor. HAMBox [32] helps outer faces compensate high-quality anchors, which can obtain high IoU regression bounding boxes. Although anchor matching has being 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 [33] adds large filters on each detection module to merge the context information. DSFD [31] introduces a feature enhance module to extend the single shot detector to dual shot detector. OS-LFFD [34] design an novel ommateum block to maintain the corresponding ratio of receptive fields to face regions. RefineFace [35] construct 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 [36], 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 various aspect ratio of face features.

#### 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 positive 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. \( B^a_i = (x^a_i, y^a_i, w^a_i, h^a_i) \) is the bounding box of anchor \( a_i \), where \( (x^a_i, y^a_i) \) is the upper-left coordinates and \( (w^a_i, h^a_i) \) 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. \( B^g_j = (x^g_j, y^g_j, w^g_j, h^g_j) \) is the bounding box of ground-truth face \( g_j \), where \( (x^g_j, y^g_j) \) is the upper-left coordinates and \( (w^g_j, h^g_j) \) represents the width and height of this face. In addition, \( r^g_j \) 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 Eq.1.

\[
\text{max IoU}(g_j) = \max_{a_i \in A} \frac{B^g_j \cap B^a_i}{B^g_j \cup B^a_i} = \max_{a_i \in A} \frac{1}{\text{Area}(B^g_j) + \text{Area}(B^a_i) - \text{Area}(B^g_j \cap B^a_i)},
\]

where \( \cap \) and \( \cup \) denote the intersection and union of two boxes respectively. \( \text{Area}(.) \) calculates the area of the bounding box. Furthermore, we recognize that there exists the max value of \( B^g_j \cap B^a_i \) formulated as in Eq.2.

\[
\text{max}(B^g_j \cap B^a_i) = \min(w^g_j, w^a_i) \times \min(h^g_j, h^a_i) = \min(w^g_j, w^a_i) \times \min(w^g_j \times r^g_j, w^a_i \times r^a_i),
\]

When anchor design determined, the width \( w^g_j \) and AR \( r^g_j \) 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 factor 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 Subsection 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 enough 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.
Table 1: The AMS is performed with different anchor aspect ratio $R^a$ and sampling threshold $T_p$. The aspect ratio sampling range of matched faces is listed below. We can approximately describe the sampling range as ARSD.

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

- 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 Eq.1 and Eq.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 Tab.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 Tab.1 of last four rows. 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 $D(r^a, \eta)$ is defined as

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

where $\eta$ is the radius of sampling domain. Thus, the left and right ARSD can be described as follows:

$$D^\times(r^a, \eta) = \{x | r^a/\eta < x < r^a \}, \quad (4)$$
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 third row of Tab.1. However, statistics show that the AR 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 attempts 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 would 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 ratio. 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.

$$E(\eta_1, \eta_0) = D(r^a, \eta_1) \backslash D(r^a, \eta_0)$$  \hspace{1cm} (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, $\eta_0$ 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.

$$E^-(\eta_1, \eta_0) = D^-(r^a, \eta_1) \backslash D^-(r^a, \eta_0)$$  \hspace{1cm} (7)

$$E^+(\eta_1, \eta_0) = D^+(r^a, \eta_1) \backslash D^+(r^a, \eta_0)$$  \hspace{1cm} (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:

$$T_p = \begin{cases} T_0 - \delta * \theta(r_j^a), & r_j^a \in E(\eta_1, \eta_0) \\ T_0, & otherwise \end{cases}$$  \hspace{1cm} (9)

$$D^+(r^a, \eta) = \{ x | r^a \leq x < r^a \eta \},$$  \hspace{1cm} (5)
where $T_0$ is the initial value of positive threshold and $\delta$ represents the amplitude of positive threshold. Similar to SAM, $T_0$ is set to 0.5. Besides, $\theta(r^g_j)$ reflects the change rate of positive threshold associated with the AR $r^g_j$ of each face. When AR of faces far away from current anchor AR, positive threshold of these face should decrease gradually. Thus, a simple implementation of $\theta(x)$ is given below.

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

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 line is our proposed WARM. Specifically, line AB and CD represent the variable positive threshold boundary of 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 $\delta$ 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.
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 no longer fit for extreme AR faces. To address this issue, we propose a novel feature enhancement module, named Receptive Field Diversity (RFD) module, to provide more diverse aspect ratio and large receptive field simultaneously. Both symmetric and asymmetric convolution kernels [36] are used to adapt the diversity of face AR.

Fig. 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 layer. Then, the 3x1, 1x3, 3x3 and 5x5 convolution kernels are employed to provide diversity receptive field. Finally, these 4-path feature maps are concatenated together. Besides, we apply a shortcut path, which maintain the original receptive field, to sum up with concatenated features above.

4 Experiments

4.1 Experimental Setup

In this section, we introduce the backbone, anchor design, data augmentation, 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 \( P_2 \),
\[
\{8\sqrt{2}, 16, 16\sqrt{2}\} \text{ in } P_3, \{32, 32\sqrt{2}, 64\} \text{ in } P_4, \{64\sqrt{2}, 128, 128\sqrt{2}\} \text{ in } P_5, \text{ and } \{256, 256\sqrt{2}, 512\} \text{ in } P_6. \text{ All anchors have aspect ratio of } 1.0.
\]

**Data Augmentation.** We randomly crop \([25, 29]\) 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 photometric color distortion \([29]\) to augment training data.

**Loss Function.** We apply the multi-task loss as our objective function. Specifically, Focal loss \([38]\) 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 \(5 \times 10^{-5}\) 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 \([39]\) 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 \([40]\) 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 Subsection 4.1 for all the experiments.
Table 2: Varying $\eta_1$, $\delta$ for WARM on WIDER FACE validation set.

| Method  | $\eta_1$ | $\delta$ | AP Easy | AP Medium | AP Hard |
|---------|----------|----------|---------|-----------|---------|
| SAM     | 2.25     | 0.00     | 0.959   | 0.950     | 0.898   |
|         | 2.50     | 0.05     | 0.962   | 0.952     | 0.899   |
| WARM    | 2.50     | 0.10     | 0.961   | 0.952     | 0.901   |
|         | 3.00     | 0.10     | 0.962   | 0.953     | 0.902   |
|         | 4.00     | 0.10     | 0.960   | 0.951     | 0.898   |

Table 3: Effectiveness of RFD module on the AP performance.

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

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 $\eta_1$, $\delta$ (defined in Subsection 3.3) is shown in Table 2. Compared with SAM, our proposed WARM can collect more positive anchors from extreme AR faces, whose ARSD is in $E(\eta_1, 2.0)$. Besides, these extra collected positive anchors have higher IoU, ranging from $0.5-\delta$ to $0.5$, related to their matched extreme AR faces. After multiple ablative experiments, we find the optimal hyper-parameters. When $\eta_1$ and $\delta$ 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 comparison experiment between SSH and RFD module as shown in Tab. 3. Previous feature enhancement modules utilize square convolutional kernels 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. From the Tab. 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 Tab. 3.
4.4 Evaluation on Benchmark

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

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 SRN [30], PyramidBox [25], FANet [22], SFDet [11], FARPN [27], ZCC [23], S*FD [29], SSH [33], HR [30], MSCNN [12], CMS-RCNN [43], MTCNN [13], LDCF+ [14], Faceness [15], Multiscale Cascade CNN [11], ACF [15], and Two-stage CNN [11]. The precision-recall curves and AP values on WIDER FACE validation and testing sets are shown in Fig. 4. It can be seen that our proposed method consistently achieves the best performance across all the three subsets. It achieves the promising average precision in all level faces, i.e. 0.965 (Easy), 0.955 (Medium), and 0.904 (Hard) for validation set, and 0.960 (Easy), 0.952 (Medium), and 0.900 (Hard) for testing set. The result in Fig. 4 demonstrates the effectiveness of our proposed method.

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 [22],
Fig. 5: Evaluation on the FDDB dataset.

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.

PyramidBox [25], DSFD [31], FD-CNN [46], ICC-CNN [47], RSA [48], S3FD [29], FaceBoxes [26], HR [20], HR-ER [20], DeepIR [49], LDCF+ [44], UnitBox [21], Conv3D [50], Faster RCNN [31], and MTCNN [19] on FDDB dataset. For a more fair comparison, the predicted bounding boxes are converted to bounding ellipses. Fig. 5 shows the discrete ROC curves and continuous ROC curves of these 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 good generalization capability of our proposed method.

4.5 Qualitative Results

Fig. 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 (see the last row of Fig. 6). Surprisingly, our detection model can capture some extra extreme AR faces which are missing labels.

Fig. 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, 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.

Fig. 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

In this paper, we examine the failure of sampling positive anchors from extreme AR faces and identify that the max IoUs of these faces are still lower than fixed sampling threshold in SAM strategy. Motivated by this observation, both Wide Aspect Ratio Matching strategies and Receptive Field Diversity module are deployed in our method for the sake of better detecting faces with different aspect ratios. 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.

6 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.

References

1. 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
2. Wu WY, Qian C, Yang S, Wang Q (2018) Look at boundary: A boundary-aware face alignment algorithm. Proc IEEE Conf Comput Vis Pattern Recognit 2129–2138
3. Liu WY, Wen YD, Yu ZD, Li M, Raj B, Song L (2017) SphereFace: Deep hypersphere embedding for face recognition. Proc IEEE Conf Comput Vis Pattern Recognit 6738–6746
4. Wang H, Wang YT, Zhou Z, Ji X, Gong DH, Zhou JC (2018) CosFace: Large margin cosine loss for deep face recognition. Proc IEEE Conf Comput Vis Pattern Recognit 5265–5274
5. Deng JK, Guo J, Xue NN, Zafeiriou S (2019) ArcFace: Additive angular margin loss for deep face recognition. Proc IEEE Conf Comput Vis Pattern Recognit 4685–4694
6. Simonyan K and Zisserman A (2014) Very deep convolutional networks for largescale image recognition. arXiv:1409.1556
7. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. Proc IEEE Conf Comput Vis Pattern Recognit 770–778
8. Huang G, Liu Z, van der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. Proc IEEE Conf Comput Vis Pattern Recognit 4700–4708
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. 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
11. Yang S, Luo P, Loy CC, Tang X (2016) Wider face: a face detection benchmark. Proc IEEE Conf Comput Vis Pattern Recognit 5525–5533
12. Jain V, Learned-Miller E (2010) FDDDB: A benchmark for face detection in unconstrained settings. University of Massachusetts. Amherst Tech Rep UM-CS-2010-009 2(7):8
13. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vis 57(2):137–154
14. 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
15. Liao S, Jain AK, Li SZ (2016) A fast and accurate unconstrained face detector. IEEE Trans Pattern Anal Mach Intell 38(2):211–223
16. Sawat DD, Hegadi RS, Garg L, Garg L (2020) Pixel encoding for unconstrained face detection. Multimed Tools Appl 79(47):35033–35054
17. Li H, Lin Z, Shen X, Brandt J, Hua G (2015) A convolutional neural network cascade for face detection. Computer Vision and Pattern Recognition IEEE 57(2):643–650
18. 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
19. 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
20. Hu P, Ramanan D (2017) Finding tiny faces. Proc IEEE Conf Comput Vis Pattern Recognit 951–959
21. 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
22. Zhang JL, Wu XW, Zhu JK, Hoi SCH (2020) Feature agglomeration networks for single-stage face detection. Neurocomputing 380:180–189
23. Liu Y, Tang X (2020) BFBox: Searching face-appropriate backbone and feature pyramid network for robust face detector. Proc IEEE Conf Comput Vis Pattern Recognit 13568–13577
24. Zhu C, Tao R, Luu K, Savvides M (2018) Seeing small faces from robust anchor’s perspective. Proc IEEE Conf Comput Vis Pattern Recognit 5127–5136
25. 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
26. 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
27. Najibi M, Singh B, Davis LS (2019) FA-RPN: Floating region proposals for face detection. Proc IEEE Conf Comput Vis Pattern Recognit 7715–7724
28. Ming X, Wei FY, Zhang T, Chen D, Wen F (2019) Group sampling for scale invariant face detection. Proc IEEE Conf Comput Vis Pattern Recognit 3441–3451
29. 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
30. 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):8231-8238
31. Li J, Wang Y, Wang C, Tai Y, Qiang J, Yang J, Wang C, Li J, Huang F (2019) Dsfd: dual shot face detector. Proc IEEE Conf Comput Vis Pattern Recognit 5055–5064
32. Liu Y, Tang X, Han JY, Liu JT, Rui D, Wu X (2020) HAMBox: delving into mining high-quality anchors on face detection. Proc IEEE Conf Comput Vis Pattern Recognit
33. 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
34. 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
35. Zhang S, Chi C, Lei Z, Li SZ (2020) RefineFace: refinement neural network for high performance face detection. IEEE Trans Pattern Anal Mach Intell
36. Ding XH, Guo YC, Ding GG, Han JG (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
37. Szegedy C, Liu W, Jia Y et al (2015) Going deeper with convolutions. Proc IEEE Conf Comput Vis Pattern Recognit 1–9
38. Lin TY, Goyal P, Girshick R, He KM, Dollar P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988
39. Shrivastava A, Gupta A, Girshick R (2016) Training region-based object detectors with online hard example mining. Proc IEEE Conf Comput Vis Pattern Recognit 761–769
40. 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
41. Zhang SF, Wen LY, Shi HL, 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
42. Cai Z, Fan Q, Feris R, 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
43. 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, Chan, pp 57–79
44. 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
45. Yang B, Yan J, Lei Z, Li S, Z. (2014) Aggregate channel features for multi-view face detection. In: Proceedings of IEEE international joint conference on biometrics, pp 1–8
46. Triantafyllidou D, Nousi P, Tefas A (2018) Fast deep convolutional face detection in the wild exploiting hard sample mining. Big Data Research 11:65–76
47. 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
48. 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
49. Sun X, Wu P, Hoi SCH (2018) Face detection using deep learning: an improved faster RCNN approach. Neurocomputing 299:42–50
50. 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
51. 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