HAMBox: Delving into Mining High-quality Anchors on Face Detection

Yang Liu*†
NCEPU
Beijing
gxly1314@gmail.com

Xu Tang* Junyu Han
Jingtuo Liu Dinger Rui
Baidu Inc.
Beijing
{tangxu02, hanjunyu,
liujingtuo, dengerrui}@baidu.com

Xiang Wu
CAS
Beijing
alfredxiangwu@gmail.com

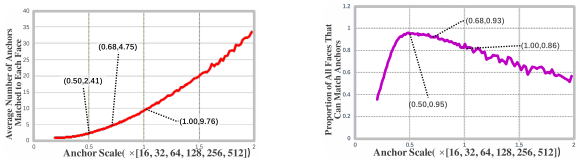
Abstract

Current face detectors utilize anchors to frame a multi-task learning problem which combines classification and bounding box regression. Effective anchor design and anchor matching strategy enable face detectors to localize faces under large pose and scale variations. However, we observe that more than 80% correctly predicted bounding boxes are regressed from the unmatched anchors (the IoUs between anchors and target faces are lower than a threshold) in the inference phase. It indicates that these unmatched anchors perform excellent regression ability, but the existing methods neglect to learn from them. In this paper, we propose an Online High-quality Anchor Mining Strategy (HAMBox), which explicitly helps outer faces compensate with high-quality anchors. Our proposed HAMBox method could be a general strategy for anchor-based single-stage face detection. Experiments on various datasets, including WIDER FACE, FDDB, AFW and PASCAL Face, demonstrate the superiority of the proposed method.

1. Introduction

Face detection is a fundamental task for many high-level face-based applications, such as face alignment [28], face recognition [1] and face aging [25]. Deriving from early face detectors with hand-crafted features, modern detectors have been significantly improved owing to the robust features learnt with deep Convolutional Neural Networks (CNNs) [10]. Current state-of-the-art face detectors are usually based on anchor-based deep CNNs, inspired by their successes on the general object detection.

Different from general object detectors, face detectors often face smaller variations of aspect ratios (from 1:1 to 1:1.5) but much larger scale variations (face area, from several pixels to thousands of pixels). Considering the large variations of scales, Zhang et al. [29] tile anchors on a wide range of layers and design anchor scales according to the effective receptive field. Current state-of-the-art detectors [21, 11] capture the locations of various face scales by utilizing Feature Pyramid Network (FPN) [12]. FPN is an effective way to exploit the inherent multi-scale features for constructing feature pyramids in a top-down manner. It adopts lateral connection from the high-level deeper features to the low-level ones. Then from the perspective of designing anchor setting, anchor-based detectors with FPN continue to resolve this by raising the number of anchors from different aspects (e.g., anchor stride, and ratio anchors [30, 23]). However, increasing the number of anchors remarkably reduces the performance of a face detector, especially when adopting the feature map conv2 or P2 (in Resnet-50) for recalling small faces empirically.



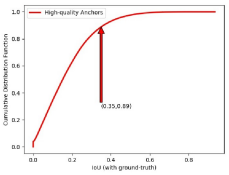
(a) Average Number of Anchors Matched to Each Face

(b) Proportion of Faces that can Match with Anchors

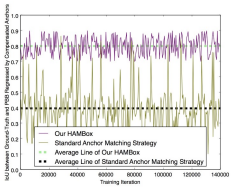
Figure 1. Two crucial factors in designing anchor scales on the WIDER FACE dataset. (a) As the scale of anchor increases, the average number of anchors matched to each face also increases. (b) The proportion of faces that can match the anchors decreases significantly outside a specific interval ([0.43, 0.7]).

As far as we know, for an anchor-based detector, effective anchor design strategies are necessary to achieve high performance. S^3FD [29] adopts single scale and aspect ratio anchors for each detection stage. Nonetheless, choosing the proper anchor scale remains a big challenge, which generally produced by the following misalignment phenomenon. Figure 1 shows 'the average number of anchors matched to each face' and 'the proportion of all faces that can match the anchors' across different anchor scales, which are two in-

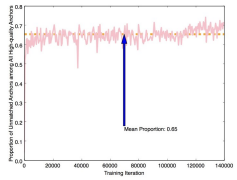
*Equal contribution. †Corresponding Author.



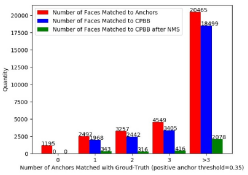
(a) Cumulative Desity Curve of IoU



(b) Performance of Compensated Anchors



(c) Proportion of unmatched High-quality Anchors



(d) Performance of Matched High-quality Anchors

Figure 2. The problem of standard anchor matching strategy during training and inference (on the WIDER FACE dataset). (a) During inference, only 11% of all correctly predicted bounding boxes are regressed by matched anchors. (b) PBB represents 'Predicted Bounding Boxes'. When using our HAMBox strategy, the IoUs between ground-truths and predicted bounding boxes regressed by compensated anchors are much higher thanstandard anchor matching strategy during training. (c) During training, the average number of unmatched high-quality anchors occupies a surprisingly 65% proportion of all high-quality anchors. (d) CPBB represents 'Correctly Predicted Bounding Boxes'. During inference, the number of matched high-quality anchors dramatically decreases after NMS, representing some unmatched anchors have higher regression ability. All these results demonstrate that the standard anchor matching strategy can not utilize high-quality negative anchors effectively, which play essential roles whatever during training or inference.

dicative factors to be considered in designing proper anchor scale. With the increase of anchor scales, although the number of anchors matched with each face steadily grows, the proportion of faces which are capable of matching anchors gradually descends. Moreover, this misalignment usually leads to a heuristical anchor scale designation.

To alleviate the imbalance between 'the average number of anchors matched to each face' and 'the proportion of all faces that can match the anchors' as discussed above, two representative solutions have been proposed: Firstly, S^{3}FD [29] introduces an anchor compensation strategy by offsetting anchors for outer faces[1]; Secondly, Zhu et al. [30] formulate a metric named Expected Maximum Overlap (EMO) to obtain more reasonable anchor stride and receptive field. All these solutions focus on helping outer faces match more anchors during the training phase. However, they also bring a large number of redundant or low-quality anchors. (see the olive line of Figure 2(b)).

In this paper, we conduct an anchor matching statistic experiment on a well-trained face detector [21] and find an intriguing phenomenon. The red line in Figure 2(a) represents the cumulative distribution curve of IoU between the ground-truth and the anchors which can be regressed to correctly predicted bounding boxes. We surprisingly observe that only 11% of all correctly predicted bounding boxes are regressed by matched anchors. So, not only the matched anchors but also some unmatched ones play a critical role in face detection. However, in the phase of training, those unmatched anchors are assigned with background labels, which are unreasonable supervision signals for classification branch consequently. Effectively leveraging these unmatched anchors is expected to improve the detection performance.

Motivated by this observation, we identify two key issues in current anchor matching strategies as follows:

- **The majority of compensated anchors are of low-quality.** Figure 2(b) shows the regression ability of compensated anchors during training when adopting the standard anchor matching strategy [18]. Apparently, compensated anchors have a poor performance on location regression (average IoU between the bounding boxes regressed by compensated anchors and the ground-truth is 0.42). In other words, this method helps those outer faces matching more low-quality anchors, instead of high-quality ones.

- **Many unmatched anchors in the training phase actually have strong localization ability**. As shown in Figure 2(c), around 65% of all high-quality anchors[2] are unmatched anchors during training. Based on the above observations, we argue that the current anchor matching strategy is neither flexible nor sufficient to utilize the anchors in face detection. As illustrated in Figure 2(d), the red, blue and green bars denote the number of faces matched to anchors (IoU$>$0.35[3]), matched to correctly predicted bounding boxes (IoU$>$0.5[4]) and matched to correctly predicted bounding boxes after Non-Maximum Suppression (NMS). It is obvious that the correctly predicted bounding boxes regressed by unmatched anchors suppress the ones regressed by matched anchors during NMS. Lots of unmatched anchors also have strong abilities for regression.

[1]Faces cannot match enough positive anchors. In our paper, we set the number as hyper-parameter K detailed in the Subsection 3.2.

[2]The intersection-over-union (IoU) between its regression bounding box and corresponding ground-truth is higher than 0.5.

[3]This denotes the IoU between anchor and target face in the training phase.

[4]This denotes the IoU between the predicted bounding box of matched anchor and the target face.

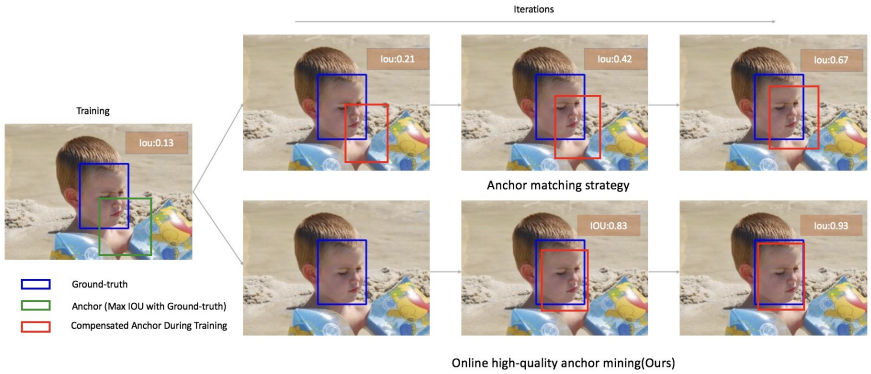


Figure 3. Visualization of the quality of compensated anchors through two methods. In the early stage of training, our method does not compensate anchors for outer faces. Then with the increasing of training iteration, our method is gradually mining unmatched high-quality anchors for outer ones, which have higher IoU than anchors generated by standard anchor matching strategy.

To address this issue, we propose an Online High-quality Anchor Mining Strategy (HAMBox) method. The idea is to mine those high-quality anchors consistently to help outer faces compensate more anchors with the ability of precise regression. Figure 2(b) and Figure 3 show that the quality of our compensated anchors has a significant enhancement than standard anchor mathing strategy's. In Figure 3, when using standard anchor matching strategy, the unmatched anchors are assigned with background labels. With the increase of training iteration, our Online High-quality Anchor Compensation Strategy is gradually mining unmatched high-quality anchors for outer faces. Moreover, the unmatched anchors could regress high-quality bounding boxes with higher IoU than anchors generated by standard anchor matching strategy. After mining high-quality anchors, we further propose regression-aware focal loss to effectively weight those new compensated high-quality anchors. Dynamic weights based on IoU are added for new compensated anchors mainly by considering the weak connection between location and classification. Benefiting from online high-quality anchor compensation strategy and regression-aware focal loss, we achieve 91.6% AP on the WIDER FACE [27] validation hard set, with the baseline of RetinaNet [13]. Furthermore, we add some popular modules, including SSH head [17], deep head [13], and pyramid anchors [21], and achieve 93.3% AP, which outperform current state-of-the-art model [11] by a large margin of 2.9% AP.

In summary, our main contributions can be summarized as:

- We observe an inspiring phenomenon that some unmatched anchors have strong regression ability, and the current box regression branch neglects to learn unmatched anchors.
- Based on the observations, we propose an Online High-quality Anchor Mining Strategy (HAMBox) to sample high-quality anchors for training. Benefiting from HAMBox, we can provide sufficient and effective anchors for outer faces in the training phase;
- Thanks to the high-quality anchors, a regression-aware focal loss assists in face detector training with a flexible way;
- Our approach outperforms the state-of-the-art methods by 2.9% and 2.3% AP on the WIDER FACE validation and test hard-set, respectively. Moreover, we achieve 57.45% (validation) / 57.13% (test) mAP on the Face Detection track of WIDER Face and Pedestrian Challenge 2019.

2. Related Work

Face detection is a fundamental yet challenging computer vision task. Viola and Jones [22] first utilizes Haar features and AdaBoost to train a face detector. After that, more following works pay attention to combining multi models to get discriminative features. For example, DPM [3] proposes an extra model to capture human lateral feature and merges it with front and back body features. All the face detectors based on hand-craft features are optimized with each sub-model separately. Due to both weak features and classifiers, the performance of these face detectors is limited in the practical scenario.

Recently, owing to the rapid development of deep convolutional networks [10, 5, 19, 20] on image classification and object detection, face detection has made significant progress on large variations, including poses, scales, blur and occlusions, etc., in practice. By introducing the core ideas of hand-craft face detector, Cascade CNN and Multi-task CNN (MTCNN) propose a coarse-to-fine framework to capture faces via deep CNNs. With the flourish of general object detectors [18, 14], [9] and SSH [17] introduce anchor-based detectors to face detection. Yang et al. [27]

collect WIDER FACE dataset, which contains rich annotations, including occlusions, poses, event categories, and face bounding boxes. The WIDER FACE dataset pushes forward to the research of face detection, focusing on the extreme variations, including scale, pose and occlusion. Recently, most state-of-the-art face detectors focus on these extreme variations with three following aspects: image pyramid, feature pyramid and context module. HR [7] designs image pyramids of the low, medium and high resolutions for training and testing, which significantly boosts the performance on extreme scale variations (from several pixels to thousands of pixels). FAN [23] introduces attention modules and feature pyramid network [12] to capture occluded faces. SSH [17] builds a detection module with a rich receptive field. PyramidBox [21] formulates a data-anchor-sampling strategy to increase the proportion of small faces in the training data. Moreover, by designing a scale propose network, SAFD [4] generates a scale histogram and further automatically normalizes face scales prior for optimizing face detectors. DSFD [11] introduces small faces supervision signals on the backbone, which implicitly boosts the performance of pyramid features.

Considering some works on anchor design and sampling strategies, S^3FD [29] proposes a new anchor matching strategy which helps the outer faces match more anchors. SRN [2] introduces a Selective Two-step Classification to ignore training easy sample anchors in the second stage. ZCC [30] introduces Expected Max Score to evaluate the quality of anchor matching, which helps to design anchor stride. Group sampling [16] conducts lots of experiments on the ratio of matched and unmatched anchors, which emphasizes the importance of the ratio for matched and unmatched anchors. In this paper, inspired by the anchor matching strategy in S^3FD [29] and the statistical curve discussed in Figure 1 and Figure 2, we propose an Online High-quality Anchor Mining Strategy (HAMBox), as well as a regression aware focal loss. Benefiting from these methods, we achieve a strong face detector, compared with other state-of-the-art face detection methods.

3. Online High-quality Anchor Mining

This section presents the proposed Online High-quality Anchor Mining Strategy (HAMBox) to compensate outer faces with the most proper anchors. We firstly build our high-recall face detector based on RetinaNet [13]. Then we demonstrate the online high-quality anchor compensation strategy in detail. Finally, we formulate a regression-aware focal loss for the compensated anchors.

3.1. High-recall Anchor-based Face Detector

Current anchor-based face detectors utilize predefined anchors to frame a multi-task learning problem, which combines classification and bounding box regression branches. We start with RetinaNet [13] as the baseline. The backbone is ResNet-50. Following the settings in [29], we employ the feature map of conv2 layer to improve the performance of face detector. The reason is that around 40% faces are matched to conv2 anchors on the WIDER FACE benchmark. Furthermore, it is important to design anchors for training a well-performed detector. Therefore, different from the general object detection with multiple anchor scales and aspect ratios, we set only one anchor scale and one aspect ratio at each prediction layer for our default anchor settings.

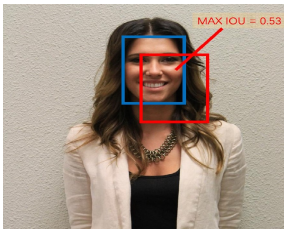
Inspired by statistical results in Figure 1, we change the anchor scale[5] to match more extreme face scales. The advantage and disadvantage of this anchor setting are equally obvious. From the perspective of advantage, our strategy can match over 95% of all the faces on the WIDER FACE benchmark, a small difference comparing to multi-scale and ratio anchors that can match 98.46% of faces. At the same time, our method uses three times or nine times fewer anchors than the latter anchor setting with multi-scale and ratio, leading model to focus more on the regression of useful anchors and further get higher detecting performance. From the perspective of disadvantage, it is harmful to the robustness of the model because decreasing the number of faces matched to anchors. This obstruction will be resolved in the following two sections.

3.2. Online High-quality Anchor Compensation Strategy

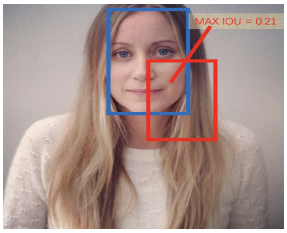
After finishing the design of anchor scale and ratio, we further need to allocate anchors with their nearest adjacent ground-truth or background. As shown in Figure 4, the current anchor matching strategy consists of two steps. A face firstly matches anchors with IoU higher than a threshold. Then faces that do not match with any anchor would be compensated with anchors that have the max IoU with them. Obviously, compensated anchors in the second step may reduce the performance of regression and classification of the network since these anchors initially have lower IoU with faces, as shown in Figure 4(c).

In Figure 2(b), we surprisingly find that with the increase of iterations, some unmatched anchors have the ability to make correct predictions while those are ignored on regression branch and even assigned as background on classification branch. Inspired by this observation, we propose an Online High-quality Anchor Compensation strategy to resolve current misaligned supervision signal. Firstly, each face matches the anchors with IoU higher than a threshold, but for those remaining outer faces, we do not compensate any anchors. Secondly, at the end of forward propagation during training, each anchor computes regression bounding

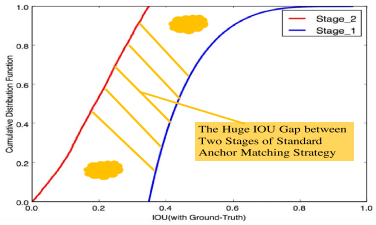
[5] In our method, anchor scale is set to $0.68 \times \{16, 32, 64, 128, 256, 512\}$ and ratio is 1:1 at different prediction layers.



(a) Face Matched to Anchor on the First Step of Standard Anchor Matching Strategy



(b) Face Matched to Anchor on the Second Step of Standard Anchor Matching Strategy



(c) Cummulative Density Curve of IoU

Figure 4. (a) (b) Two different stages on standard anchor matching strategy, the blue rectangle represents ground-truth and the red one is an anchor matched with it. (c) Cumulative Density Curve of IoU between ground-truth and its matched anchor on different stages.

box through its related regression coordinates. We define this regression bounding box as B_{reg} and F_{outer} represents outer faces. Finally, for each face in F_{outer}, we compute its IoU with B_{reg} and compensate this face with N extra unmatched anchors. We define all IoUs as IoU_{set}. These N compensated anchors are selected according to two rules. 1) The IoUs between their corresponding regression bounding boxes and target faces should be greater than T (T represents an online positive anchor threshold). 2) These IoUs (calculated in rule 1) should be in the top-K highest IoU in IoU_{set}. K is a hyperparameter that represents the max number of anchors that F_{outer} can be matched with. If N is greater than $K - M$ after filtering out by above two rules, we select top-$(K - M)$ highest IoU anchors in these N unmatched anchors to compensate this face and set $N = K - M$. M denotes the number of anchors that faces already matched with in the first step. We have done many experiments in ablation study by varying T, K. Details can be seen in Algorithm 1.

3.3. Regression-aware Focal Loss

After the analysis of two subsections above, we have mined those high-quality anchors and the following problem is to make full use of these anchors effectively. Furthermore, we propose a regression-aware focal loss to give more reasonable weights to those new compensated high-quality anchors, which are newly mined for outer faces by Online High-quality Anchor Compensation Strategy.

Two improvements have been made on focal loss [13]. (1) Considering the weak connection between location and classification on new compensated anchors, we add dynamic weights based on IoU for these compensated ones. (2) We define anchors satisfying the following three conditions simultaneously as ignored anchors (which are not optimized during training): a) Belong to the high-quality anchors. b) Be assigned as background in the first step of anchor matching strategy. c) Not included in new compensated anchors. We define the loss as:

$$\begin{aligned} L_{cls}(p_i) &= \frac{1}{N_{com}} \sum_{i \in \psi} F_i L_{fl}(p_i, g_i^*) \\ &+ \frac{1}{N_{norm}} \sum_{i \in \Omega} (1_{(l_i^*=0)} 1_{(F_i<0.5)} + 1_{(l_i^*=1)}) L_{fl}(p_i, l_i^*) \end{aligned} \tag{1}$$

where i is the anchor index in a training-batch, p_i is the predicted probability of the anchor i. ${l_i}^*$ is the class label of anchor i, which is assigned with the label on the first step of standard anchor matching strategy. ${g_i}^*$ is the label of our newly compensated anchors, which are revised from backgrounds to foregrounds. F_i is the IoU between the corresponding regression bounding box and its target ground-truth. Ω represents a set of all matched and unmatched low-quality anchors[6] and ψ represents a set of newly compensated anchors. N_{norm} is the number of normally matched anchors in Ω and N_{com} is the total number of compensated anchors in ψ. L_{fl} is the normal sigmoid focal loss over two classes (face foreground and background). In addition, the supervision for new compensated anchors is added to the location loss and the specific equation is shown as below:

$$\begin{aligned} L_{loc}(x_i) &= \frac{1}{N_{com}} \sum_{i \in \psi} L_{SmoothL1}(x_i, x_i^*) \\ &+ \frac{1}{N_{norm}} \sum_{i \in \Omega} L_{SmoothL1}(x_i, x_i^*) \end{aligned} \tag{2}$$

where ${x_i}^*$ is the ground-truth location coordinates of anchor i. $L_{SmoothL1}$ is a normal location loss inspired by Faster-RCNN [18]. All other parameters are similar to L_{cls}'s.

4. Experiments

In this section, we first show the effectiveness of our proposed strategies with comprehensive ablative experiments. Then with the final optimal model, our approach achieves state-of-the-art results on face detection benchmarks.

[6]This denotes the IoU between the predicted bounding box regressed by unmatched anchor and the target face is below 0.5.

Algorithm 1 Online high-quality anchor mining

Input: B, G, T, K, D, L, R, A

B is a set of regression bounding boxes, in the form of (x_0, y_0, x_1, y_1).

X is a set of ground-truth, in the form of (x_0, y_0, x_1, y_1)

T is an online anchor mining threshold (see details on Subsection 3.2)

K is a hyperparameter and represents the max number of anchors that F_{outer} can be matched with.

D is a Dict, key is ground-truth, item is the number of anchors that ground-truth can match in the first step of our HAMBox method.

L is a Dict, key is anchor index, item is a label that anchor index is assigned with in the final process of our HAMBox method.

R is a Dict, key is anchor index, item is encoded coordinates of the key during standard anchor matching strategy.

A is a Dict, key is anchor index, item is coordinates of the key, in the form of (x_0, y_0, x_1, y_1).

Output: R and L after using our HAMBox method.

for x_i in X **do**
  if $D(x_i) >= K$ **then**
    $continue$
  end if
  $CompensatedNumber = K - D(g_i)$
  $OnlineIoU \Leftarrow IoU(x_i, B), AnchorIdx$
  $SortedOnlineIoU = sorted(OnlineIoU, key = IoU, reverse = True, Iou > T)$
  if $len(SortedOnlineIoU) = 0$ **then**
    $continue$
  end if
  for $IoU, AnchorIdx$ in $SortedOnlineIoU$ **do**
    if $L(AnchorIdx) = 1$ **then**
      $continue$
    end if
    $CompensatedNumber$ -= 1
    $L(AnchorIdx) = 1$
    $R(AnchorIdx) = encode(A(AnchorIdx), ground\text{-}truth)$
    if $CompensatedNumber = 0$ **then**
      $break$
    end if
  end for
end for
Return R, L

4.1. Ablation Study

The WIDER FACE dataset is used in this ablation study. This dataset has 32,203 images with 393,703 labeled faces with huge variability in scales, occlusions and poses. Our networks are only trained on the training set and evaluated both on validation and test set. Average Precision (AP) score is used as the evaluation metric.

Data Augmentation Our models are trained with following data augmentation strategies:

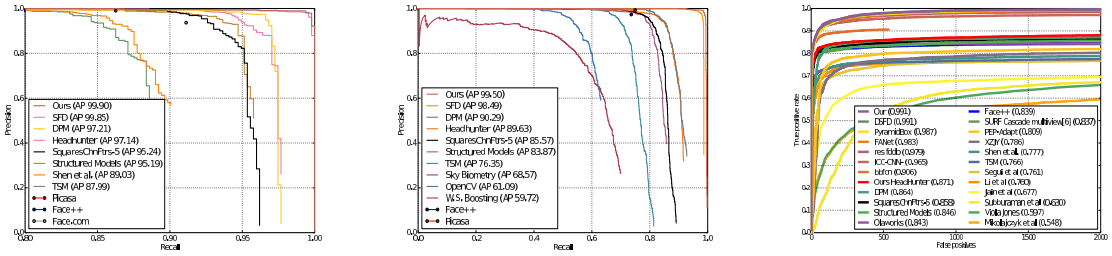
- Color distort: Apply some photometric distortions similar to [6].
- Data anchor sampling: This method [21] resizes all train images through reshaping a random face in this image to a smaller size.
- Horizontal flip: After data-anchor-sampling, the cropped image patch is resized to 640 $\times$ 640 and horizontally flipped with a probability of 0.5.

Baseline. We build an anchor-based detector with ResNet-50 guided by the RetinaNet as our baseline face detector. It differs from the original RetinaNet [13] in the following four aspects: Firstly, we set 6 anchors whose scales are from the set {16, 32, 64, 128, 256, 512}, and all anchors' aspect ratios are set to 1:1. Secondly, we use LFPN [21] instead of FPN [12] for feature fusion since top two high-level features are extracted from regions with little context and may introduce noise for detecting small faces. Thirdly, we do not use deep head owing to two main factors. One is that the time cost of the training process is too high, the other is that our baseline is significantly higher than that of any other SOTA works on the WIDER FACE hard dataset. Finally, the threshold of IoU for matched anchors is changed to 0.35, and ignore-zone is not implemented.

Optimization Details. All models are initialized with the pre-trained weights of ResNet-50 and fine-tuned on WIDER FACE training set. Each training iteration contains seven images per GPU for a 4 NVIDIA Tesla V100 GPUs server. The initial learning rate is set to 0.01 and decreases to 0.001 after 110k iterations. All the models are trained for 140k iterations by synchronized SGD. The momentum and weight decay are set to 0.9 and 5×10^{-5}, respectively.

The effect of High-recall Anchor-based Detector As discussed above, the difference between our high-recall detector and baseline detector is the pre-defined anchor scale. Inspired by Figure 1, we design our method with anchor scales {16, 32, 64, 128, 256, 512} * 0.68 tiled on pyramid feature maps from P2 to P6.

To better understand the advantage of our method, we conduct four experiments, which are shown in Table 1. First step on standard anchor matching strategy with anchor scale ratio 0.68 (denoted as SMS(ratio=0.68)); Two-step on standard anchor matching strategy with anchor scale ratio 0.68 (denoted as DMS(ratio=0.68)); Two-step on standard anchor matching strategy with anchor scale ratio 0.5 whose scale could help more outer faces match anchor while significantly decreasing the number of anchors matched



(a) AFW (b) PASCAL Face (c) FDDB

Figure 5. Evaluation on the common face detection datasets.

with each face (denoted as DMS(ratio=0.50)); New anchor matching strategy introduced by S^{3}FD [29] with anchor scale ratio 0.68 (denoted as NAMS(ratio=0.68)). Compared to the baseline, SMS(ratio=0.68), DMS(ratio=0.68), NAMS(ratio=0.68) provide a significant improvement on the hard subset (rising by 1.2%, 0.8%, 0.8% AP respectively) and DMS(ratio=0.50) is with no improvements on the hard subset (decreasing by 0.7%). Through these experimental results, we could draw two conclusions: On the one hand, enhancing the proportion of faces that can be matched with anchors could improve the model performance. However, with the continuously decrease on the scale of anchor to enhance this proportion, the remaining faces are more difficult to match and the number of anchors that each face can match with decreases dramatically, which are the main reasons why the performance of DMS(ratio=0.50) is 1.5% AP lower than DMS(ratio=0.68) on the hard dataset. On the other hand, NAMS(ratio=0.68) and DMS(ratio=0.68) achieve almost same performance with SMS(ratio=0.68), suggesting that these two anchor compensation methods have less influence on the performance of the detector. Thus we use the SMS(ratio=0.68) method in the following experiments. In addition, DMS(ratio=0.68) and NAMS(ratio=0.68) would be regarded as comparisons, respectively. And the anchor ratio of SMS in Table 3 and 4 is set to 0.68.

Table 1. AP performance on various anchor setting and anchor matching strategy on WIDER FACE validation subset.

Subset	ratio	Easy	Medium	Hard
Baseline	1.00	0.943	0.931	0.894
+SMS	0.68	**0.949**	**0.945**	**0.906**
+DMS	0.68	0.954	0.949	0.902
+DMS	0.50	0.938	0.922	0.887
+NAMS	0.68	0.951	0.948	0.902

The Effect of Online High-quality Anchor Mining Next, we look into the effect of our proposed online high-quality anchor mining strategy. In this paragraph, we mainly discuss the effect of two hyper-parameters in our method. The performance under different K, T (defined in Subsection 3.2) is shown in Table 2. It shows that: (1) the performance gets better when T increases and it is easier to conclude that the higher the quality of compensated anchors, the better the performance of the model. (2) The performance gets better when K is smaller than 5 and gets worse when K is larger than 5, suggesting that it is not good to increase too large numbers of compensated anchors that each outer face can match since the anchors off-limits are redundant for their corresponding faces. After multiple ablative experiments, we find the optimal K(3), T(8) and further increase the performance with 0.7% AP.

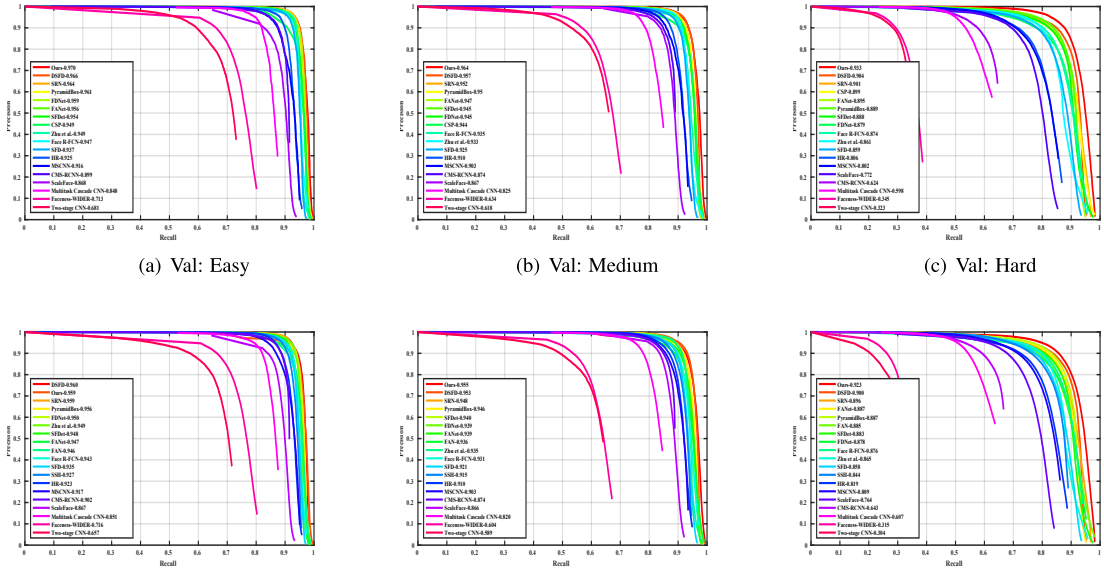
Table 2. Varying T, K for regression-aware focal loss on WIDER FACE validation subset.

K	T	Easy	Medium	Hard
3	0.5	0.945	0.939	0.902
7	0.5	0.941	0.937	0.898
3	0.7	0.947	0.943	0.911
5	0.7	0.952	0.941	0.909
3	**0.8**	**0.957**	**0.951**	**0.913**
3	0.9	0.962	0.943	0.911

The Effect of Regression-aware Focal Loss This regression-aware focal loss completes our final HAMBox model. As discussed above, this loss gives those new matched anchors a reasonable weight which simultaneously helps model training these anchors more steadily and precisely. Results using our regression-aware focal loss (denoted as RAL) are shown in Table 4, and the performance of our detector continues to increase 0.3% AP.

Our method vs ZCC and NAMS To further verify the effectiveness of our method, we compare our method with NAMS [29] and ZCC [30]. As shown in Table 3, our method outperforms theirs 6.4% and 5.5% AP respectively on their paper baseline. Moreover, in our baseline, ours also outperforms theirs 2.0% and 3.4% AP, respectively. Note that our method offers more high-quality anchors to help bounding box regression and classification branch optimize well.

The Effect of Other Tricks As shown in Table 4, we introduce SSH [17], deep head (DH) [13] and pyramid anchor (PA) [21] modules to further improve the performance of detector and achieve best AP among all state-of-the-art face detectors [11, 2, 15, 21, 24, 30, 29, 17, 7].



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

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

Figure 6. Precision-Recall (PR) curves on WIDER FACE validation and testing subsets.

Table 3. AP performance of our model with various anchor matching strategy on WIDER FACE validation subset. * denotes the reproduced performance by us, and APs in NAMS and ZCC represent the performance presented by their papers.

Subset	Easy	Medium	Hard
NAMS	0.937	0.924	0.852
ZCC	0.949	0.933	0.861
Baseline + NAMS*	0.941	0.937	0.896
Baseline + ZCC*	0.943	0.942	0.882
Baseline + SMS + OAM + RAL	**0.962**	**0.953**	**0.916**

Table 4. AP performance of our proposed modules and additional tricks on WIDER FACE validation subset.

Baseline	SMS	OAM	RAL	DH	SSH	PA	Easy	Medium	Hard
√	-	-	-	-	-	-	0.943	0.931	0.894
√	√	-	-	-	-	-	0.949	0.945	0.906
√	√	√	-	-	-	-	0.957	0.951	0.913
√	√	√	√	-	-	-	0.962	0.953	0.916
√	√	√	√	√	-	-	0.964	0.955	0.922
√	√	√	√	√	√	-	0.968	0.959	0.927
√	√	√	√	√	√	√	**0.970**	**0.964**	**0.933**

4.2. Evaluation on Common Benchmarks

We evaluate our proposed method on the common face detection benchmarks, including WIDER FACE [27], Annotated Faces in the Wild (AFW) [31], PASCAL Faces [26], FDDB [8]. Our face detector is trained only using WIDER FACE training set and is tested on those benchmarks.

WIDER FACE Dataset We report the performance of our face detection system on the WIDER FACE [27] testing set with 16,097 images. Detection results are sent to the database server for receiving the precision-recall curves. Figure 6 illustrates the precision-recall curves along with AP scores. Our proposed method achieves 97.0% (Easy), 96.4%(Medium), 93.3%(Hard) on validation dataset and 95.9% (Easy), 95.5% (Medium), 92.3% (Hard) on test dataset.

AFW Dataset This dataset [31] consists of 205 images with 473 annotated faces. Figure 5(a) shows that our detector outperforms others by a considerable margin.

PASCAL Face Dataset This dataset [26] has 851 images with 1,335 annotated faces. Figure 5(b) demonstrates the superiority of our method.

FDDB Dataset This dataset [8] has 2,845 images with 5,171 annotated faces. Most of them are with low image resolutions and complicated scenes, such as occlusions, huge poses. Figure 5(c) shows our proposed method outperforms all state-of-the-art models.

5. Conclusion

In this paper, we first observe an interesting phenomenon that only 11% correctly predicted bounding boxes are regressed from the unmatched anchors in the inference phase. Then we further propose an online high-quality anchor mining strategy that helps outer faces match high-quality anchors. Our method first enhances the proportion of face matched with anchor, and then we propose an online high-quality anchor compensation strategy for outer faces. Finally, we design a regression-aware focal loss for new compensated anchors. We conduct extensive experiments on the AFW, PASCAL Face, FDDB, WIDER FACE datasets and achieve the state-of-the-art detection performance.

References

[1] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks. In *ICIP*, pages 2089–2093. IEEE, 2017.

[2] Cheng Chi, Shifeng Zhang, Junliang Xing, Zhen Lei, Stan Z Li, and Xudong Zou. Selective refinement network for high performance face detection. In *AAAI*, volume 33, pages 8231–8238, 2019.

[3] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2009.

[4] Zekun Hao, Yu Liu, Hongwei Qin, Junjie Yan, Xiu Li, and Xiaolin Hu. Scale-aware face detection. In *CVPR*, pages 6186–6195, 2017.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[6] Andrew G Howard. Some improvements on deep convolutional neural network based image classification. *arXiv preprint arXiv:1312.5402*, 2013.

[7] Peiyun Hu and Deva Ramanan. Finding tiny faces. In *CVPR*, pages 951–959, 2017.

[8] Vidit Jain and Erik Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. 2010.

[9] Huaizu Jiang and Erik Learned-Miller. Face detection with the faster r-cnn. In *FG*, pages 650–657. IEEE, 2017.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[11] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: dual shot face detector. In *CVPR*, pages 5060–5069, 2019.

[12] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017.

[13] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

[14] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016.

[15] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. High-level semantic feature detection: A new perspective for pedestrian detection. In *CVPR*, pages 5187–5196, 2019.

[16] Xiang Ming, Fangyun Wei, Ting Zhang, Dong Chen, and Fang Wen. Group sampling for scale invariant face detection. In *CVPR*, pages 3446–3456, 2019.

[17] Mahyar Najibi, Pouya Samangouei, Rama Chellappa, and Larry S Davis. Ssh: Single stage headless face detector. In *ICCV*, pages 4875–4884, 2017.

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[20] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[21] Xu Tang, Daniel K Du, Zeqiang He, and Jingtuo Liu. Pyramidbox: A context-assisted single shot face detector. In *ECCV*, pages 797–813, 2018.

[22] Paul Viola and Michael J Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.

[23] Jianfeng Wang, Ye Yuan, and Gang Yu. Face attention network: An effective face detector for the occluded faces. *arXiv preprint arXiv:1711.07246*, 2017.

[24] Yitong Wang, Xing Ji, Zheng Zhou, Hao Wang, and Zhifeng Li. Detecting faces using region-based fully convolutional networks. *arXiv preprint arXiv:1709.05256*, 2017.

[25] Zongwei Wang, Xu Tang, Weixin Luo, and Shenghua Gao. Face aging with identity-preserved conditional generative adversarial networks. In *CVPR*, pages 7939–7947, 2018.

[26] Junjie Yan, Xuzong Zhang, Zhen Lei, and Stan Z Li. Face detection by structural models. *Image and Vision Computing*, 32(10):790–799, 2014.

[27] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *CVPR*, pages 5525–5533, 2016.

[28] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *ECCV*, pages 1–16. Springer, 2014.

[29] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *ICCV*, pages 192–201, 2017.

[30] Chenchen Zhu, Ran Tao, Khoa Luu, and Marios Savvides. Seeing small faces from robust anchor's perspective. In *ICCV*, pages 5127–5136, 2018.

[31] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886. IEEE, 2012.