MogFace: Towards a Deeper Appreciation on Face Detection

Yang Liu [1] Fei Wang [1] Jiankang Deng [2] Zhipeng Zhou [1] Baigui Sun [1] Hao Li [*1]
[1]Alibaba Group [2]Imperial College London

Abstract

Benefiting from the pioneering design of generic object detectors, significant achievements have been made in the field of face detection. Typically, the architectures of the backbone, feature pyramid layer, and detection head module within the face detector all assimilate the excellent experience from general object detectors. However, several effective methods, including label assignment and scale-level data augmentation strategy, fail to maintain consistent superiority when applying on the face detector directly. Concretely, the former strategy involves a vast body of hyper-parameters and the latter one suffers from the challenge of scale distribution bias between different detection tasks, which both limit their generalization abilities. Furthermore, in order to provide accurate face bounding boxes for facial down-stream tasks, the face detector imperatively requires the elimination of false alarms. As a result, practical solutions on label assignment, scale-level data augmentation, and reducing false alarms are necessary for advancing face detectors. In this paper, we focus on resolving three aforementioned challenges that exiting methods are difficult to finish off and present a novel face detector, termed MogFace. In our Mogface, three key components, Adaptive Online Incremental Anchor Mining Strategy, Selective Scale Enhancement Strategy and Hierarchical Context-Aware Module, are separately proposed to boost the performance of face detectors. Finally, to the best of our knowledge, our MogFace is the best face detector on the Wider Face leader-board, achieving all champions across different testing scenarios. The code is available at https://github.com/damo-cv/MogFace.

1. Introduction

Face detector, predicting location coordinates of face boxes, serves as the fundamental step for many facial down-stream tasks, including face alignment [1], face recognition [5] [28] and face attribute analysis [24]. In the past few years, we have witnessed the quick development on the general object detectors, deriving from the Fast-RCNN [8] and SSD [18] to Retinanet [16] and DERT [3]. Motivated by this, state-of-the-art face detectors adopt the great architecture designs from general object detectors, such as Feature Pyramid Network [15] and One-stage Single-Shot framework [16].

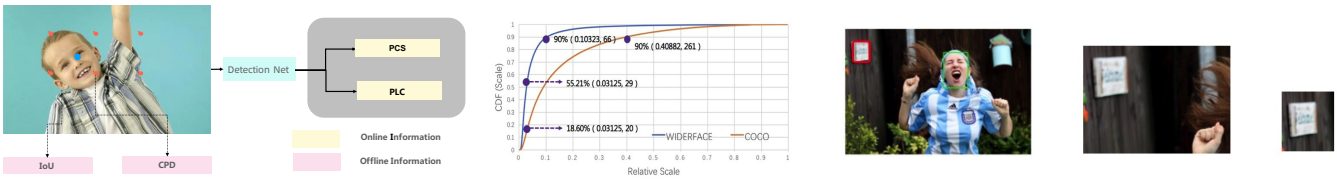
However, label assignment and scale-level data augmentation strategy [1], achieving great superiority on the task of generic object detection, bring rare gains on face detectors . On the one hand , the designation of former strategy involves a vast body of hyper-parameters (e.g. K in ATSS [34], α in OTA [7]) , which limits its generalization ability. On the other hand, as shown in Fig. 1(b), compared with generic object detector, face detector confronts more severe scale variance challenge. Uniform sampling based scale-level data augmentation strategies (e.g. multi-scale training and random square crop [18]), serving as the main scale enhancement methods on generic object detectors [16, 23, 34], fail to provide effective scale information for face detector (more analysis can be seen in the supplementary material). Furthermore, the face detector is a real-world application that emphasizes reducing the number of false alarms urgently. Therefore, how to distinguish false alarms away from true positive faces is an another distinctive challenge on the task of face detection.

Based on the above analysis, we discover that label assignment strategy, scale-level data augmentation strategy and eliminating false alarms have a huge potential for constructing a high-performance face detector. Then we perform a systematically quantitative and qualitative analysis on 3 aforementioned perspectives to provide some intrinsic insights.

Label Assignment. Label assignment strategies adopt predefined rules to match ground-truth (gt) or background for each anchor. As shown in the Fig. 1(a), the designation of predefined rules highly depends on offline and online information. Offline information contains Intersection-over-Union (IoU) and Center Point Distance (CPD) between gt

* Corresponding Author.
Email: ly261666@alibaba-inc.com

[1]enriches the scale distribution of the training data to resolve scale variance challenge.



(a) (b) (c)

Figure 1: Motivation illustration. (a) Online and offline information both can be adopted as criterion to determine the boundary between positive and negative anchors. But how to effectively and adaptively combine them remains a huge challenge. (b) Cumulative density curve of face or object scale relative to the fixed scale (640). In the Wider face and COCO dataset, almost 55% and 18% ground-truth scale is less than 20, demonstrating that compared to generic object detector, a more severe scale variance challenge is occurred on the task of face detection. (c) For the same detector, we discover the top-left calendar is a false alarm in the left image, while the top-left calendar in the other two images are not.

and anchor, which can be computed during the process of data preparation. Online information consists of the predicted classification scores (PCS) and the predicted location coordinates (PLC), which can be extracted at the end of forward propagation. Traditional label assignment strategies adopt offline information as threshold criterion for pos/neg anchors division, e.g. IoU in retinanet [16], faster-rcnn [23], IoU and CPD in ATSS [34]. Recently, Hambox [21] further points out the effectiveness of online information and put forwards an online high-quality anchor mining strategy to utilize the PLC. OTA [7] formulates the assigning procedure as an optimal transport problem, where the cost function is designed by the weighed combination of CPD, PCS and PLC.

However, there exist two drawbacks lying behind current label assignment strategies: 1) Online information cannot provide high-confidence matching information as well as offline one. Thus, it will result in the emergence of sub-optimal label assignment strategy when encouraging online rules to serve as main metric on distinguishing positive and negative anchors like OTA and Hambox. 2) The selection of hyper-parameter in the recent label assignment strategies frequently goes through constant trials and errors, making it difficult for transferring different detection tasks, e.g. from general object detection to face detection. In this paper, we address the aforementioned issues by proposing an adaptive online incremental anchor mining strategy (Ali-AMS), which is based on the standard anchor matching strategy adopted in retinanet [16] and further compensates outlier ground-truths with incremental anchors at the end of forward propagation. In our Ali-AMS, two key components, quality assessment based anchor mining strategy and pyramid-level consistency principle, are proposed to mine and assign the high-quality anchor adaptively. Concretely, the former strategy regards the PCS as the quality assessment criterion to re-sort the anchor mined with the CPD and IoU information; then, the latter principle guarantees that ground-truths located at the same pyramid layer can match the same number of anchors. The motivation and more details of our Ali-AMS can be seen in section 3.1.

Scale-level Data Augmentation. Generic object detector frequently introduces scale-level data augmentation strategy to resolve extreme scale variance on the COCO dataset. However, the most authoritative face detection dataset Wider Face [31] contains more severe scale variance than COCO [17]. As shown in Fig. 1(b), we display the scale distribution of ground-truths on the COCO and Wider Face benchmark, respectively. Comparing with COCO dataset which is famous with extreme scale variance, Wider Face has more rigorous scale distribution, where contains almost 55% small scale faces [2].

To resolve extreme scale variance challenge, there exist 3 widely-adopted data augmentation strategies, including Multi-scale-training (MST), Random Square Crop (RSP) and Data-anchor-sampling (DAS). Multi-scale-training strategy, resizing each image into a random scale selected from fixed scale range, frequently serves as the optimal solution on handling with scale variance problem, which has demonstrated its significance in many technology reports on the COCO detection challenge. Random Square Crop strategy, cropping the square area from a given image randomly, is a main-stream scale-level data augmentation strategy on the task of face detection [6], [35]. Data-anchor-sampling strategy [26] aims to introduce more small scale faces by resizing each image into a smaller scale.

MST and RSP are both designed from uniform sampling perspective while DAS focuses on generating many small faces. Meanwhile, in the Wider Face training set, almost 55% face scale is less than 20, making uniform sampling based augmentation strategies generate a large number of small faces. As a result, MST and RSP both have great detection ability on small faces. This raises a worth solving problem: How to increase the detection ability on the

[2] the scale of ground-truth is less than 20

middle and large scale faces since the training set only contains a small proportion large faces (10%) ? In this paper, we investigate this question by analyzing the relationship between the performance of each pyramid layer and the number of ground-truths it matches. Based on comprehensive quantitative and qualitative analysis in supplementary material, we unexpected discover that it is not the more ground-truths that is matched in a single pyramid layer, the greater performance of this pyramid layer. As a result, this phenomenon releases an amazing conclusion that in order to improve the representation of certain pyramid layer, the number of ground-truths matched in this layer should be appropriate instead of 'the more, the better'. Under the guidance of this meaningful conclusion, we propose a simplex selective scale enhancement strategy for detecting large scale face accurately on the basis of prior statistical result, which controls the ground-truths distribution to improve the deep pyramid layer representation, achieving the best detection performance on large and middle scale faces. To the best of our knowledge, our selective scale enhancement strategy is the first novel work to consider the relationship between the performance of each pyramid layer and the number of ground-truths it matches, which provides a solid knowledge on how to mine the learning capacity of certain pyramid layer.

Eliminating False Alarms. Reducing the number of false alarms is vitally important for the real-world face detector. The common solution is to introduce additional training data with false alarms, which helps the detector acquire more knowledge on the property of false alarms. However, collecting extra training data is labor-intensive such that the solution without extra data is deserved exploring. In this paper, we present a Hierarchical Context-Aware Module (HCAM) to help false alarms away from ground-truths, which explicitly encodes neighbour context information into high-confidence anchors [3]. The effectiveness of neighbour context information can be seen in Fig. 1(c), we send the left image into the Hambox [21] face detector and find a top-left false alarm. However, when we crop this false alarm with expanding context area (middle image) and less expanding context area (right image), we unexpectedly discover that the same detector believes there is no false positives in these two images. Moreover, we utilize Hambox face detector to find all false alarms in the Wider Face validation dataset, almost 95% false alarms are disappeared when adopting similar operations like above. This phenomenon demonstrates that the appropriate neighbour context information is conducive to eliminating false alarms.

In summary, our contribution can be summarized as:

- Presenting 3 worthy of in-depth research topics on the task of face detection, including Label Assignment, Scale-level Data Augmentation and Eliminating False Alarms.
- Proposing Adaptive Online Incremental Anchor Mining Strategy, Selective Scale Enhancement Strategy, Hierarchical Context-Aware module respectively to construct a promising face detector, termed as MogFace.
- Achieving state-of-the-art results in all popular face detection benchmarks, including Wider Face, AFW, FDDB and Pascal Face.

[3]high-confidence anchors contains correct predicted positive anchors and false predicted negative anchors

2. Related Work

Recently, a large number of face detectors [6, 9, 13, 19, 20,26] has been proposed to advance face detection community. In this section, we mainly review the related work from two following perspectives, label assignment and scale-level data augmentation strategies.

Label Assignment. In the field of face detection, zhang et.al. [35] propose a scale compensation anchor matching strategy to increase the matched anchors of outer faces by reducing the IoU threshold. Liu et al. [21] discover that some negative anchors have stronger localization ability than positive ones and further propose an "online high-quality anchor mining strategy" by compensating outer faces with high-quality negative anchors. In the field of object detection, ATSS [34] automatically selects positive and negative samples according to statistical characteristics of object. OTA [7] formulates the label assigning procedure as an optimal transport problem, which converts the best assignment solution into solving the optimal transport plan at minimal transportation costs.

Scale-level Data Augmentation. SNIP [25] introduces a Scale Normalization for Image Pyramids (SNIP) strategy which selectively back-propagates the gradients of object instances of different sizes as a function of the image scale. Li et.al [14] proposes a novel Trident Network to generate scale-specific feature maps with a uniform representational power. Data-anchor-sampling [26] augments the training samples to increase the diversity of training data for smaller faces. Zhang et.al [35] proposes a scale-equitable face detection framework to handle different scales of faces. ASFD [32] introduces an automatic feature enhance module to allow multi-scale feature fusion efficiently.

3. Method

In this section, we consecutively introduce three components of our MogFace, including Adaptive Online Incremental Anchor Mining Strategy, Selective Scale Enhancement Strategy and Hierarchical Context-Aware Module.

Algorithm 1 Adaptive Online Incremental Anchor Mining Strategy (Ali-AMS)

Input:

$\mathcal{A}$ is a dict, key is ground-truth, value is the number of anchors matched with this ground-truth with standard anchor matching strategy.

1: **for** pyramid layer pi in $[p2, p3, p4, p5, p6, p7]$ **do**
2: $\mathcal{G} \leftarrow$ ground-truths that matched in the pi
3: $\mathcal{T} \leftarrow$ maximum number of anchors that matched in $\mathcal{G}$ with standard anchor matching strategy.
4: **for** each ground-truth $g \in \mathcal{G}$ **do**
5: **if** $\mathcal{A}[g] == \mathcal{T}$ **then**
6: continue
7: **end if**
8: compute the number of compensated anchors for g: $\mathcal{N}_g = \mathcal{T} - \mathcal{A}[g]$;
9: $cpd \leftarrow$ select $\mathcal{T}$ anchors whose centers are closest to the center of ground-truth g based on L2 distance;
10: $iou \leftarrow$ select $\mathcal{T}$ anchors whose have the highest iou with ground-truth;
11: $conf\_candidate \leftarrow$ sort the anchor in iou and cpd according to predicted classification scores;
12: select top-$\mathcal{N}_g$ confident anchors from $conf\_condidate$ to serve as positive anchors for g;
13: **end for**
14: **end for**

3.1. Adaptive Online Incremental Anchor Mining Strategy

Algorithm 1 describes how our Ali-AMS compensates high-quality anchors for outer ground-truths, when given an input image. Concretely, our Ali-AMS consists of pyramid-level consistency principle and quality assessment based anchor mining strategy, which identifies the number of anchors matched with each ground-truth among all pyramid layers and mines the high-quality anchors for outer ground-truth, respectively. The first principle (line 2-3) is described as follows: on each pyramid level, we first find the set of ground-truths ($\mathcal{G}$) matched in this pyramid layer based on the standard anchor matching strategy [16] and compute the maximum number ($\mathcal{T}$) of anchors that matched in $\mathcal{G}$. For each gt (g) in $\mathcal{G}$, if the number of anchors matched with g is less than $\mathcal{T}$, the number of anchors that the g matches will be compensated to T with following metric. The description of second strategy is from line 8 to 12: For each ground-truth that needs to be matched with incremental anchors, the select process of compensated anchors contains three following steps: 1) select $\mathcal{T}$ anchors from the view of CPD and IoU separately. 2) Sort all 2 * $\mathcal{T}$ anchors by the predicted classification score 3) Compensate top-$\mathcal{N}_g$ anchors for the outer ground-truth. Note that the number of top-$\mathcal{N}_g$ is computed in the first principle adaptively. The motivation behind our Ali-AMS is explained as follows.

Adopting predicted classification score as quality assessment on the candidates mined with CPD and IoU information. We explain this from the view on the necessity and the role of online information separately. 1) The success of anchor-based detector demonstrates that only utilizing offline information (CPD or IoU) as metric to determine the boundary between pos/neg anchors can provide a great optimization direction for the detector. However, with the number of iterations increasing, offline information can not provide the progressive matching rules, resulting in the sub-optimal optimization direction. Analogously, Hambox [21] points out the same conclusion that even though adopting offline information as metric to distinguish pos/neg anchros, many negative anchors have amazing regression ability. Such inconsistent phenomenon between the learnt knowledge on detector and the formulated knowledge on offline information (IoU) suggests that offline information based label assignment strategy fails to satisfy the requirements of optimization process. Therefore, combing offline and online information can provide more accurate optimization direction, due to online information can reflect the learnt knowledge on the detector. 2) Relying too much on the online rule to identify pos/neg anchors has an obvious drawback that it is easily over fitting on the bias assignment strategy, especially on the early stage of optimization process. Thus, rather than serving the online information as dominant criterion, we regard online rule as quality assessment tool (auxiliary role) to measure the candidates which already have mined with CPD and IoU rules.

3.2. Selective Scale Enhancement Strategy

As discussed above, previous scale-level data augmentation strategies fail to resolve the challenging problem: How to increase the detection ability on the medium and large scale faces since the training set only contains a small proportion large faces (10%) ? To resolve this, we first analyze the relationship between the performance of each pyramid layer and the number of ground-truths it matches in the supplementary material. To our surprise, we discover an amazing conclusion: it is not accurate that the more ground-truths matched in one pyramid layer, the greater performance of this pyramid layer. Based on this instructive finding, we propose a selective scale enhancement strategy (SSE) to maximize the learning capacity of deeper pyramid layer by controlling the distribution of ground-truths among pyramid layers from $p2$ to $p7$ based on the prior statistical result.

To begin with, we introduce three prerequisites for our SSE strategy. (1) We firstly define the conception of the main pyramid layer and the auxiliary pyramid layer, which embraces the top-2 greatest detection abilities on the large-scale faces among all pyramid layers according to the empirical results reported in the supplementary material. As

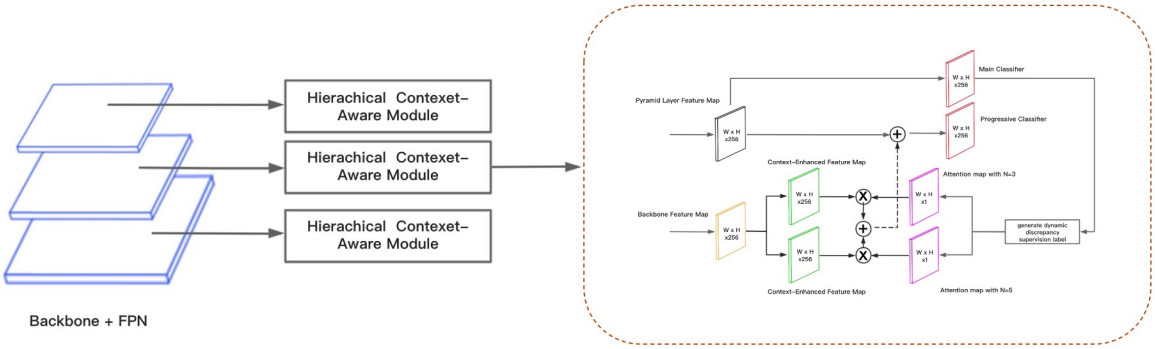


Figure 2: Hierarchical Context-Aware Module.

a result, p5/p6 are the main/auxiliary pyramid layer for the SSE strategy. (2) Then we determine the ratio of the ground-truths matched in the main and auxiliary pyramid layer. Let r_pi (i=2,3,4,5,6,7) represents the maximum performance ratio matched in the pyramid layer pi. The maximum performance ratio refers to that this ratio together with our proposed scale control strategy [4] achieves the best performance among all candidate ratios, that is shown on the table 1 with bold annotation. Thus, r_p5 equals to 20% and r_p6 equals to 20%. For the main pyramid layer, we define the ratio of the total ground-truths matched in the main pyramid layer as tr_mpl that equals to r_p5. Moreover, we define the ratio of the total ground-truths matched in the auxiliary pyramid layer as tr_apl that equals to (1 - tr_mpl) $*$ r_p6. This assignment strategy on the scale information of ground-truths can guarantee the learning capacity of the main and auxiliary pyramid layer successively. (3) Finally, in the table 2, we compute the scale range (sr_pi, i=2,3,4,5,6,7) of the faces that matched in different pyramid layers. Note that the overlap scale range between neighbour pyramid layers is divided uniformly for convenience.

	20%	40%	60%	80%
p4 (easy)	67.0	75.3	81.3	**84.3**
p4 (med)	75.3	82.6	**84.5**	83.6
p5 (easy)	81.8	**82.2**	77.9	73.2
p5 (med)	**85.4**	85.2	85.0	83.7
p6 (easy)	**86.2**	83.1	84.8	83.3
p6 (med)	**85.2**	80.5	82.2	81.2

Table 1: The results of scale control strategy on the Wider Face validation subsets.

Based on 3 above prerequisites, algorithm 2 shows the pipeline of our SSE strategy on each training image. 1) Resize image by reshaping the short side of image into a scale selected from scale range [640, 1280]. 2) Randomly sample a face from the resized image and compute its scale fs. 3) Identify the target pyramid layer (tpl) . Randomly sample a floating-point number (rn) from 0 to 1. If the value of the rn is less than tr_p5, we define the target pyramid layer as $p5$. If rn is over r_p5 and less than r_p5 + r_p6, the target pyramid layer (tpl) equals to p6. If rn is over r_p5 + r_p6, the target pyramid layer equals to the random one of the pyramid layers except for $p5$ and $p6$. 4) Random select a scale from sr_tpl and compute the target resize ratio (trr) by fs / this scale. 5) Resize image with trr. 6) Define the resolution of the input image as N $\times$ N. If the resolution of the resized image is over N $\times$ N, we crop N $\times$ N area randomly as the input image and pad zero pixel if it is less than N $\times$ N. The motivation behind our SSE strategy is as follows.

[4]Details shown in the supplementary materials

	start_scale	end_scale
sr_p2	8.4	20.7
sr_p3	20.7	48.2
sr_p4	48.2	106.2
sr_p5	106.2	212.4
sr_p6	212.4	420.8
sr_p7	420.8	640

Table 2: The scale range of the ground-truths that matched in different pyramid layers.

Depending on the main and auxiliary pyramid layer to control the scale distribution of ground-truths. The previous designation on controlling the scale distribution of the ground-truths can be divided into following two types. MST and RCS emphasize on the uniform sampling while DAS points out an alternative view that the more small-scale data can boost the detection ability on small faces. However, such heuristic designation can not utilize the scale information effectively. On the one hand, in the supplementary

Algorithm 2 Selective Scale Enhancement Strategy (SSE)

Input:

$\mathcal{I}$ is a set of all training data

tr_p5 is a hyperparameter that represents the ratio of the total ground-truths matched in the main pyramid layer (p5).

tr_p6 is a hyperparameter that represents the ratio of the total ground-truths matched in the auxiliary pyramid layer (p6).

sr_tpl is the scale range of the ground-truths that matched in the tpl.

$\mathcal{N}$ is the side of the resolution for the input image.

Output:

$\mathcal{S}$ is a set of training data augmented by our SSE strategy

for each image $i \in \mathcal{I}$ **do**
  $\mathcal{R}_i \leftarrow$ Resize image by reshaping the short side of image into a scale selected from scale range [640, 1280] randomly.
  $\mathcal{F}_s \leftarrow$ Compute the face scale that is selected from $\mathcal{R}_i$ randomly.
  $Random\_float = Random.random(0, 1)$
  if $Random\_float < tr\_p5$ **then**
    $tpl = p5$
  else if $Random\_float <= (tr\_p5 + tr\_p6)$ **then**
    $tpl = p6$
  else
    $tpl = random([p2, p3, p4, p7])$
  end if
  compute target resize ratio: $trr = random(sr\_tpl)/\mathcal{F}_s$;
  $\mathcal{R}_i^{trr} \leftarrow$ Resize $\mathcal{R}_i$ with the shrink ratio trr;
  if $resolution(\mathcal{R}_i^{trr}) > (\mathcal{N}, \mathcal{N})$ **then**
    $\mathcal{R}_i^{trr} \leftarrow$ crop $\mathcal{N} \times \mathcal{N}$ area randomly from $\mathcal{R}_i^{trr}$;
  else
    $\mathcal{R}_i^{trr} \leftarrow$ expand $\mathcal{R}_i^{trr}$ into $(\mathcal{N}, \mathcal{N})$ by padding zero pixel;
  end if
  if $\mathcal{S}$ is None **then**
    $\mathcal{S} \leftarrow \varnothing$
  end if
  $\mathcal{S} = \mathcal{S} \cup \mathcal{R}_i^{trr}$;
end for
return $\mathcal{S}$;

material, we find RSC and DAS achieve almost consistent performance on the Wider Face hard subset although DAS brings more small faces. On the other hand, as described in the table 1, we find it is not accurate that the more ground-truths that is matched in one pyramid layer, the greater performance of this pyramid layer. These two findings demonstrate that heuristically controlling the scale distribution of ground-truths fail to satisfy the ground-truths requirements on the related pyramid layer. Therefore, our SSE, consecutively satisfying the requirement of main and auxiliary pyramid layer on the ground-truths, is a better strategy than other heuristic designation.

3.3. Hierarchical Context-Aware Module

As analyzed above, we find the appropriate neighbour context information is conducive to eliminating false alarms. Motivated by this finding, we propose a Hierarchical Context-Aware Module (HCAM) to distinguish false alarms away from the face by encoding the neighbour context information into related high-confidence anchors in Fig. 2. Under the rescue of neighbour context information, false predicted negative anchors and correct predicted positive anchors within the high-confidence anchors can be separated explicitly. The training pipeline is as follows:

1) We firstly send a batch of images into the detector and get the classification score of each anchor on the main classifier. Then we mask the position of all high-confidence anchors to generate two attention feature maps. For each attention feature map, we assign the position of high-confidence anchors and their related neighbour information [5] as 1, otherwise is 0. The N are set 3 and 5 for generating different neighbour information.

2) We compute the two neighbour context information by dot-multiplying the context-enhanced backbone feature maps and two attention feature maps, respectively. This step aims to explicitly generate more abundant context information related to the high-confidence anchors. Note that more ablative experiments and architectures on context-enhanced modules are discussed in the supplementary material.

3) We further encode two neighbour context information into pyramid feature map by element-wise summation operator.

4) Finally, this combined feature map feeds into progressive classifier to further distinguish false predicted negative anchors (false alarms) away from correct predicted positive anchors according to Equ. 1.

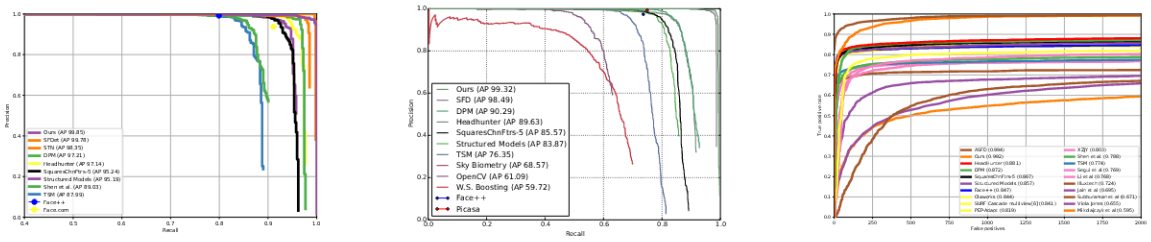
$$L = f_{fl}^{m}(f_c, y) + \gamma * f_{fl}^{p}(f_c^{com}, y_{hc}) \tag{1}$$

where f_{fl}^{m} and f_{fl}^{p} represent the sigmoid focal loss [16] over two classes, that are applied on the main and progressive classifier, respectively. y is the the label of anchor which is assigned by anchor matching strategy [23]. The weight γ aims to balance the loss between main and progressive classifiers (we find γ=1 has a best performance). fc and f_c^{com} represent the output of main classifier and progressive classifier, separately. y_{hc} is dynamic discrepancy supervision label, that is defined by following two steps: 1) In each iteration of the training stage, we firstly mask the position of correctly predicted positive anchors [6] and false predicted negative anchors [7] according to the predicted clas-

[5] neighbour information takes the target anchor point as the center and contains the surrounding N×N area, N is an integer greater than 0

[6] refer to the positive anchors which have high confidence classification score

[7] refer to the negative anchors which have high confidence classification score



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

Figure 3: Evaluation on common face detection benchmarks.

sification score on the main classifier at the end of forward propagation. 2) Assigning the position of correct predicted positive anchors as positive samples, the position of false predicted negative anchors as negative samples, the position of remaining anchors as ignore samples.

4. Experiment

In this section, we first elaborate the implementation details of our baseline. Then, we conduct ablative experiments on the most authoritative face detection benchmark, Wider Face. Finally, we test our method on all popular face datasets, including AFW, PASCAL Face, FDDB and Wider Face.

4.1. Baseline Setup

Training Detail. We adopt SFD [35] architecture with Resnet50 [10] backbone as our baseline. For anchor settings, 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. The losses applied on our baseline are focal loss and smooth L1 loss with the weight ratio 1:2. For optimization details, each training iteration contains 7 images per GPU on 4 NVIDIA Tesla V100s. Models are optimized by synchronized SGD. The momentum and weight decay are set to 0.9 and 5×10-5, respectively. For learning rate schedule, the initial learning rate is set to 4e-3 and decreases to 4e-4 in 50000 iterations. The total iteration is 70000.

Inference Detail. In the phase of inference, we adopt single scale test strategy to evaluate the experiment result. Firstly, we feed the image with original scale into the detector and then get top-5000 highest confidence bounding boxes. Then, the Non-maximum Suppression is applied with the IoU threshold 0.6 to get top 750 confident detection scores and related bounding boxes.

4.2. Ablation Study

The Effectiveness of Ali-AMS. In the table 3, we compare our Ali-AMS with recent proposed label assignment strategies, including HAMBox [21], OTA [7] and ATSS [34]. Our Ali-AMS achieves the best performance among all recent label assignment strategies on the Wider Face hard subset with outperforming baseline, HAMBox, ATSS, OTA by 0.8%, 0.4%, 1.0%, 2.6% AP score, respectively.

Method	Easy	Medium	Hard
Baseline	94.6	93.4	86.5
Baseline + HAMBox	94.5	**93.8**	86.9
Baseline + ATSS	94.5	93.2	86.3
Baseline + OTA	93.3	91.7	84.7
Baseline + Ali-AMS	**94.6**	93.6	**87.3**

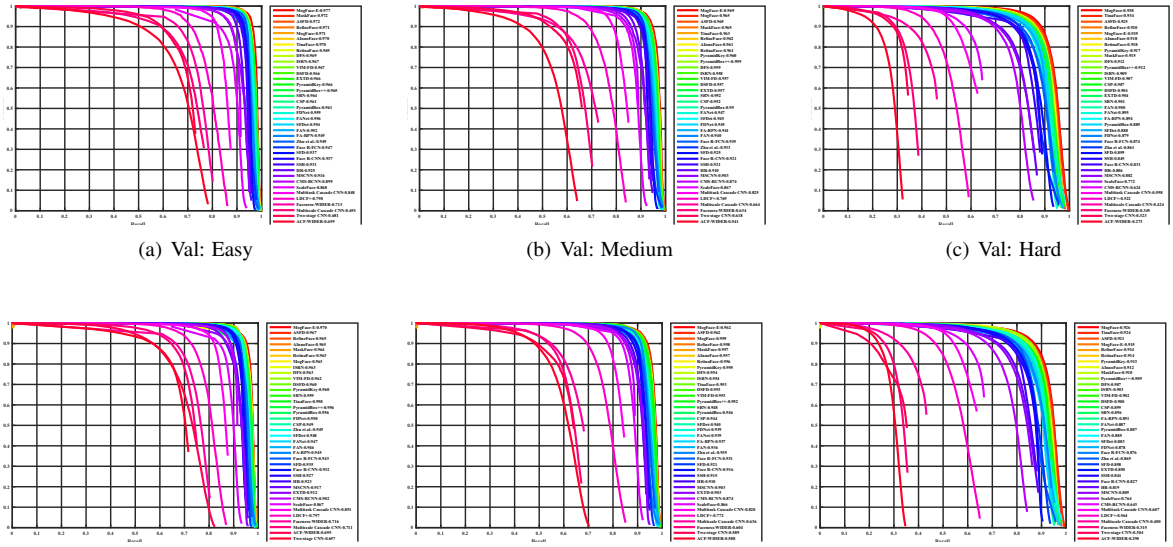
Table 3: Results of Ali-AMS on the Wider Face validation subsets.

The Effectiveness of SSE. In the table 4, we compare our SSE with other scale-level data augmentation strategies, including data-anchor-sampling, random square crop and multi scale training strategy. Our SSE outperforms others by 1.0%, 0.7% AP on the Wider Face validation easy, medium subset respectively. Such tremendous enhancement on detecting large-scale faces help our MogFace-E achieve 4 champions on the Wider Face official leaderboard.

Method	Easy	Medium	Hard
Baseline (w/o DAS)	92.2	90.5	81.4
Baseline + MST	93.3	91.5	83.6
Baseline + DAS	94.6	93.4	**86.5**
Baseline + SSE	**95.6**	**94.1**	-

Table 4: Results of SSE on the Wider Face validation subsets.

The Effectiveness of HCAM. Table 5 presents the performance and the number of false alarms (NFA) on our Hierarchical Context-Aware false alarms Module with different N (defined in the footnote 5). The Hierarchical Context-Aware Module achieves the best performance (87.4 % AP on the Wider Face Validation hard subset) and contains the minimal false alarms, when N=3 and 5. Note that the context-aware module can also only contain one neighbour context information, e.g. N=3 or N=5.



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

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

Figure 4: Precision-Recall (PR) curves on Wider Face validation and testing subsets.

Method	N	Easy	Medium	Hard	NFA
Baseline	-	94.6	93.4	86.5	948
HCAM	3	94.9	94.0	86.8	532
HCAM	5	94.8	94.1	87.0	476
HCAM	3 and 5	**95.1**	**94.2**	**87.4**	**192**

Table 5: Results of our Hierarchical Context-Aware Module on the Wider Face validation subsets.

4.3. Evaluation on Common Benchmarks

In this subsection, we compare our MogFace with state-of-the-arts methods on the common face detection benchmarks, including AFW [38], Pascal Face [30], FDDB [12] and Wider Face [31]. We train the MogFace-E (Ali-AMS, HCAM, SSE) and MogFace (Ali-AMS, HCAM) with some excellent modules introduced by Hambox [21] detector on the Wider Face training dataset, including SSH head [22], Pyramid Anchor [26] and deep head [16].

AFW Dataset. This dataset contains 205 images with 473 annotated faces. As shown in Fig. 3(a), our Mogface significantly outperforms other methods by 1.0% AP at least.

PASCAL Face Dataset. This dataset contains 851 images with 1335 annotated faces. Fig. 3(b) shows that our method achieves the state-of-the-art results by outperforming the second one with 0.8% AP.

FDDB Dataset. This dataset has 2,845 images with 5,171 annotated faces. Most of them have low image resolutions and complicated scenes . Fig. 3(c) shows that our MogFace achieves the highest (99.2%) performance.

Wider Face Dataset. We test our MogFace and MogFace-E with Multi-scale results ensemble strategy on the Wider Face validation and test set. As shown in Fig. 4, we plot a precision-recall curve according to the official tool on the Wider Face validation set. As for Wider Face test set, we submit the detection bounding boxes and corresponding scores to the official server to get the precision-recall curves that is shown in Fig. 4. Our method achieves 97.7% (Easy), 96.9% (Medium), 93.8% (Hard) AP performance on the Wider Face validation set and 97.0 % (Easy), 96.2% (Medium), 92.6% (Hard) AP performance on the Wider Face test set. Comparing with other sota approaches [2, 4, 11, 13, 27, 29, 32, 33, 36, 37], our method outperforms them by 0.5% (Validation Easy), 0.5 % (Validation Medium), 0.4 % (Validation Hard), 0.3 % (Test Easy), 0.2 % (Test Hard). Such tremendous enhancement in all scenarios demonstrates the superiority of our MogFace.

5. Conclusion

In this paper, we first point out that the success experience from generic object detection fails to provide effective solutions on label assignment, scale-level data augmentation, and eliminating face alarms in the field of face detection. Thereby, in order to advance the development of face detectors, we resolve three aforementioned challenges by proposing Adaptive Online Incremental Anchor Mining Strategy, Selective Scale Enhancement Strategy, Hierarchical Context-Aware Module, respectively. Finally, benefiting from prominent solutions of our MogFace, we achieve six champions on the Wider Face dataset, which continues to this day.

References

[1] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017. 1

[2] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *European conference on computer vision*, pages 354–370. Springer, 2016. 8

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1

[4] Cheng Chi, Shifeng Zhang, Junliang Xing, Zhen Lei, Stan Z Li, and Xudong Zou. Selective refinement network for high performance face detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8231–8238, 2019. 8

[5] J Deng, J Guo, and S Zafeiriou. Arcface: additive angular margin loss for deep face recognition. corr abs/1801.07698 (2018), 1801. 1

[6] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019. 2, 3

[7] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 303–312, 2021. 1, 2, 3, 7

[8] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1

[9] Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. *arXiv preprint arXiv:2105.04714*, 2021. 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[11] Peiyun Hu and Deva Ramanan. Finding tiny faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 951–959, 2017. 8

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

[13] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5060–5069, 2019. 3, 8

[14] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 6054–6063, 2019. 3

[15] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1

[16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1, 2, 4, 6, 8

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[18] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 1

[19] Yang Liu and Xu Tang. Bfbox: Searching face-appropriate backbone and feature pyramid network for face detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13568–13577, 2020. 3

[20] Yang Liu, Xu Tang, Junyu Han, Jingtuo Liu, Dinger Rui, and Xiang Wu. Hambox: Delving into mining high-quality anchors on face detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13043–13051. IEEE, 2020. 3

[21] Yang Liu, Xu Tang, Xiang Wu, Junyu Han, Jingtuo Liu, and Errui Ding. Hambox: Delving into online high-quality anchors mining for detecting outer faces. *arXiv preprint arXiv:1912.09231*, 2019. 2, 3, 4, 7, 8

[22] Mahyar Najibi, Pouya Samangouei, Rama Chellappa, and Larry S Davis. Ssh: Single stage headless face detector. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4875–4884, 2017. 8

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1, 2, 6

[24] Ying Shu, Yan Yan, Si Chen, Jing-Hao Xue, Chunhua Shen, and Hanzi Wang. Learning spatial-semantic relationship for facial attribute recognition with limited labeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11916–11925, 2021. 1

[25] Bharat Singh and Larry S Davis. An analysis of scale invariance in object detection snip. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3578–3587, 2018. 3

[26] Xu Tang, Daniel K Du, Zeqiang He, and Jingtuo Liu. Pyramidbox: A context-assisted single shot face detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 797–813, 2018. 2, 3, 8

[27] Jianfeng Wang, Ye Yuan, and Gang Yu. Face attention network: An effective face detector for the occluded faces. *arXiv preprint arXiv:1711.07246*, 2017. 8

[28] Kai Wang, Shuo Wang, Zhipeng Zhou, Xiaobo Wang, Xiaojiang Peng, Baigui Sun, Hao Li, and Yang You. An efficient

training approach for very large scale face recognition. *arXiv preprint arXiv:2105.10375*, 2021. 1

[29] 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. 8

[30] Junjie Yan, Xuzong Zhang, Zhen Lei, and Stan Z Li. Face detection by structural models. *Image and Vision Computing*, 32(10):790–799, 2014. 8

[31] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016. 2, 8

[32] Bin Zhang, Jian Li, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, Yili Xia, Wenjiang Pei, and Rongrong Ji. Asfd: Automatic and scalable face detector. *arXiv preprint arXiv:2003.11228*, 2020. 3, 8

[33] Changzheng Zhang, Xiang Xu, and Dandan Tu. Face detection using improved faster rcnn. *arXiv preprint arXiv:1802.02142*, 2018. 8

[34] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9759–9768, 2020. 1, 2, 3, 7

[35] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 192–201, 2017. 2, 3, 7

[36] Chenchen Zhu, Ran Tao, Khoa Luu, and Marios Savvides. Seeing small faces from robust anchor's perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5127–5136, 2018. 8

[37] Chenchen Zhu, Yutong Zheng, Khoa Luu, and Marios Savvides. Cms-rcnn: contextual multi-scale region-based cnn for unconstrained face detection. In *Deep learning for biometrics*, pages 57–79. Springer, 2017. 8

[38] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2879–2886. IEEE, 2012. 8