Seeing Small Faces from Robust Anchor's Perspective

Chenchen Zhu Ran Tao Khoa Luu Marios Savvides
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213, USA
{chenchez, rant, kluu, marioss}@andrew.cmu.edu

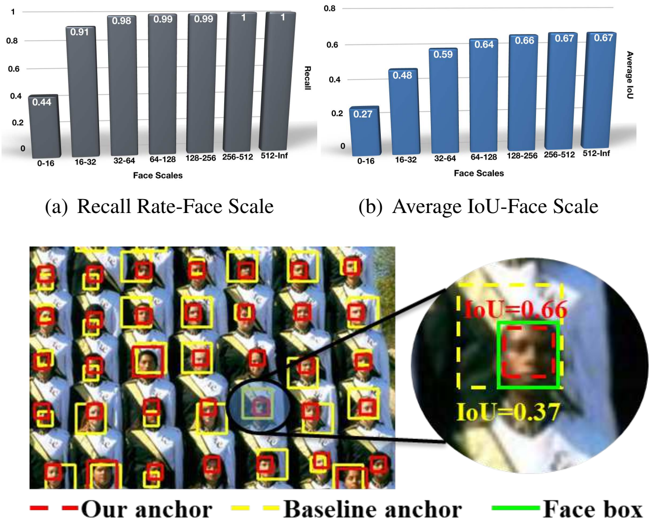
Abstract

This paper introduces a novel anchor design principle to support anchor-based face detection for superior scale-invariant performance, especially on tiny faces. To achieve this, we explicitly address the problem that anchor-based detectors drop performance drastically on faces with tiny sizes, e.g. less than 16 × 16 pixels. In this paper, we investigate why this is the case. We discover that current anchor design cannot guarantee high overlaps between tiny faces and anchor boxes, which increases the difficulty of training. The new Expected Max Overlapping (EMO) score is proposed which can theoretically explain the low overlapping issue and inspire several effective strategies of new anchor design leading to higher face overlaps, including anchor stride reduction with new network architectures, extra shifted anchors, and stochastic face shifting. Comprehensive experiments show that our proposed method significantly outperforms the baseline anchor-based detector, while consistently achieving state-of-the-art results on challenging face detection datasets with competitive runtime speed.

1. Introduction

Face detection plays an important role in many facial analysis tasks [1, 37, 21]. Starting from the traditional face detectors with hand-crafted features [36, 51, 8, 45, 16], the modern detectors [42, 17, 27, 47] have been dramatically improved thanks to the robust and discriminative features from deep convolutional neural networks (CNNs) [33, 10]. Current state-of-the-art face detectors are designed using *anchor-based* deep CNNs [50, 11, 44, 26, 48], inspired by the techniques from popular object detectors [20, 30, 4].

Anchor-based detectors quantize the continuous space of all possible face bounding boxes on the image plane into the discrete space of a set of pre-defined anchor boxes that serve as references. The Intersection-over-Union (IoU) overlap is used for evaluating the similarity between anchor boxes and face boxes. During training, each face is matched with one or several close anchors. Matched anchors are trained to output high confidence scores and then regress to ground-truth boxes. During inference, faces in a testing image are detected by classifying and regressing anchors.



(a) Recall Rate-Face Scale (b) Average IoU-Face Scale

(c) Baseline anchors (yellow) vs. our anchors (red) with higher face IoUs

Figure 1. Problems of current anchor-based detectors and our solution. **(a):** A baseline anchor-based detector, trained and evaluated on the Wider Face dataset (see Section 5.1 for details), has significantly lower recall rate at IoU of 0.5 on tiny faces (16 × 16) than larger faces. **(b):** Maximum IoU with anchors is computed for each face and averaged in each scale group, showing positive correlation with the recall rate across the scales. **(c):** Visualization of the anchor boxes with highest IoUs for each face. Our anchors have much higher IoU with faces than the baseline anchor. Right side shows an enlarged example. (Best viewed in color)

Although anchor-based detectors have shown success in handling shape and appearance invariance [6, 19], their capabilities in handling scale invariance is not satisfactory. On the other hand, faces can be captured at any size in images. In addition to heavy occlusion, extreme pose and low illumination, very small faces have become one of the most challenging problems in robust face detection. Figure 1(a) shows the face recall rate of a baseline anchor-based detector across different face scales. While big faces (larger than 64×64 pixels) can be almost 100% recovered, there is a significant drop in recall for smaller faces, especially those with less than 16×16 pixels. In other words, after classifying and adjusting anchor boxes, the new boxes with high confidence scores are still not highly overlapped with enough small faces. This suggests that we look at how anchors are overlapped with faces initially before training. For each face we compute its highest IoU with overlapped anchors. Then faces are divided into several scale groups. Within each scale group we compute the averaged highest IoU score, as presented in Figure 1(b). It's not surprising to find that average IoUs across face scales are positively correlated with the recall rates. We argue that anchors with low IoU overlaps with small faces are harder to be adjusted to the ground-truth, resulting in low recall of small faces.

In this paper, we focus on new anchor design to support anchor-based detectors for better scale-invariance. Our newly proposed anchors have higher IoU overlaps with faces than the baseline anchors, as shown in Figure 1(c). Therefore, it is easier for the network to learn how to adjust the new anchors (red boxes) to ground-truth faces than the original anchors (yellow boxes). To achieve this, we look deep into how faces are matched to anchors with various configurations and propose the new Expected Max Overlapping (EMO) score to characterize anchors' ability of achieving high IoU overlaps with faces. Specifically, given a face of known size and a set of anchors, we compute the expected max IoU of the face with the anchors, assuming the face's location is drawn from a 2D distribution of on the image plane. The EMO score theoretically explains why larger faces are easier to be highly overlapped by anchors and serves as a guidance for our anchor design principle.

The EMO score enlightens several simple but effective strategies of new anchor design for higher face IoU scores without introducing much complexity to the network. Specifically, we propose to reduce the anchor stride with various network architecture designs. We also propose to add anchors shifted away from the canonical center so that the anchor distribution becomes denser. In addition, we propose to stochastically shift the faces in order to increase the chance of getting higher IoU overlaps. Finally, we propose to match low-overlapped faces with multiple anchors.

We run extensive ablative experiments to show our proposed method can achieve significant improvement over the baseline anchor-based detector. It also achieves the state-of-the-art results on challenging face detection benchmarks, including Wider Face [43], AFW [22], PASCAL Faces [41], and FDDB [12] In summary, the main contributions of this paper are three folds: **(1)** Provide an in-depth analysis of the anchor matching mechanism under different conditions with the newly proposed Expected Max Overlap (EMO) score to theoretically characterize anchors' ability of achieving high face IoU scores. **(2)** Propose several effective techniques of new anchor design for higher IoU scores especially for tiny faces, including anchor stride reduction with new network architectures, extra shifted anchors, and stochastic face shifting. Demonstrate significant improvement over the baseline anchor-based method. **(3)** Achieve state-of-the-art performance on Wider Face, AFW, PASCAL Faces and FDDB with competitive runtime speed.

2. Related Work

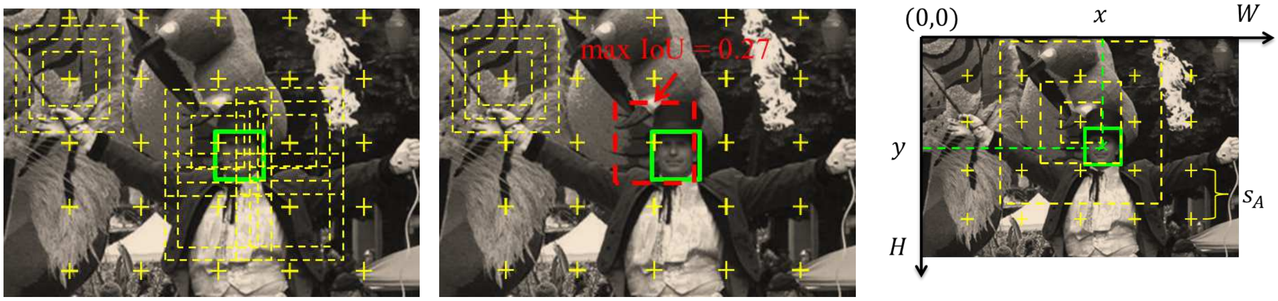
Face detection is a mature yet challenging computer vision problem. One of the first well performing approaches is the Viola-Jones face detector [36]. Its concepts of boosting and using simple rigid templates have been the basis for different approaches [46, 16]. More recent works on face detection tend to focus on using different models such as a Deformable Parts Model (DPM) [51, 8, 45, 2, 9]. Mathias et al. [25] were able to show that both DPM models and rigid template detectors have a lot of potential that has not been adequately explored. All of these detectors extract handcrafted features and optimize each component disjointly, which makes the training less optimal.

Thanks to the remarkable achievement of deep convolutional networks on image classification [33, 10] and object detection [30, 4, 20], deep learning based face detection methods have also gained much performance improvement recently [42, 17, 50, 27, 47, 44, 11, 34, 26, 48]. CNNs trained on large-scale image datasets provide more discriminative features for face detector compared to traditional hand-crafted features. The end-to-end training style promotes better optimization. The performance gap between human and artificial face detectors has been significantly closed.

However, the Wider Face dataset [43] pushes the challenge to another level. In addition to heavy occlusion, extreme pose, and strong illumination, the ultra small sizes of faces in crowd images have become one of the most challenging problems in robust face detection. To solve this, CMS-RCNN [50] incorporates body contextual information to help infer the location of faces. HR [11] builds multilevel image pyramids for multi-scale training and testing which finds upscaled tiny faces. SFD [48] addresses this with scale-equitable framework and new anchor matching strategy. In this paper, we introduce a novel anchor design for finding more tiny faces, leading to state-of-the-art detection performance.

3. Expected Max Overlapping Scores

This section presents the new Expected Max Overlapping (EMO) score to characterize anchors' ability of achieving high face IoU scores. We start with an overview of



(a) Anchor setup and distribution

(b) Anchor matching mechanism

(c) Computing the EMO score

Figure 2. **(a):** Anchors are a set of boxes with different sizes (yellow dashed boxes) tiled regularly (centered on "+" crosses) on the image plane. A face (green) overlaps with multiple anchors. **(b):** A face is matched to an anchor with the max IoU. The matched anchor is highlighted as red dashed box. **(c):** The EMO score characterizes the anchors' ability of capturing a face by computing the expected max IoU of the face with anchors w.r.t. the distribution of face's location (Best viewed in color).

anchor-based detectors. Then we show the standard anchor setup and the anchor matching mechanism. Finally we derive the EMO by computing the expected max IoU between a face and anchors w.r.t. the distribution of face's location.

3.1. Overview of Anchor-Based Detector

Anchor-based detection methods classify and regress anchor boxes to detect objects. Anchors are a set of pre-defined boxes with multiple scales and aspect ratios tiled regularly on the image plane. During training, anchors are matched to the ground-truth boxes based on the IoU overlap. An anchor will be assigned to a ground-truth box if a) its IoU with this box is the highest than other anchors, or b) its IoU is higher than a threshold T_h. An anchor will be labeled as background if its IoU overlaps with all boxes are lower than a threshold T_l.

Anchors are associated with certain feature maps which determine the location and stride of anchors. A feature map is a tensor of size $c \times h \times w$, where c is the number of channels, h and w are the height and the width respectively. It can also be interpreted as c-dimensional representations corresponding to $h \cdot w$ sliding-window locations distributed regularly on the image. The distance between adjacent locations is the feature stride s_F and decided by $\frac{H}{h} = \frac{W}{w} = s_F$. Anchors take those locations as their centers and use the corresponded representations to compute confidence scores and bounding box regression. So, the anchor stride s_A is equivalent to the feature stride, i.e. $s_A = s_F$.

3.2. Anchor Setup and Matching

We consider the standard anchor setup as shown in Figure 2. Let S be a pre-defined scale set representing the scales of anchor boxes, and R be a pre-defined ratio set representing the aspect ratios of anchor boxes. Then, the number of different boxes is $|S \times R| = |S||R|$, where $\times$ is the Cartesian product of two sets and $| * |$ is the set's cardinality. For example, anchor boxes with 3 scales and 1 ratio are shown as the yellow dashed rectangles in the top-left corner in Figure 2(a). Let L be the set of regularly distributed locations shown as the yellow crosses "+", with the distance between two adjacent locations as anchor stride s_A. Then the set of all anchors A is constructed by repeatedly tiling anchor boxes centered on those locations, i.e. $A = S \times R \times L$.

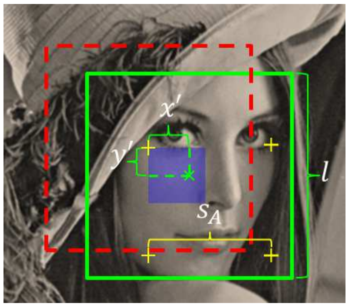


Figure 3. The EMO score is a function of face scale l and anchor stride s_A (Best viewed in color).

Given a face box B_f shown as the green rectangle, it is matched to an anchor box B_a with the max IoU overlap shown as the dashed red rectangle (Figure 2(b)). The max IoU overlap is computed as Eq. (1).

$$\max_{a \in A} \frac{|B_f \cap B_a|}{|B_f \cup B_a|} \tag{1}$$

where $\cap$ and $\cup$ denote the intersection and union of two boxes respectively.

3.3. Computing the EMO Score

A face can randomly appear at any location on the image plane $W \times H$, where H and W are the height and width respectively. In Figure 2(c), we denote the center point of a face as a pair of random variables (x, y), i.e. the green cross "$\times$". Let $p(x, y)$ be the probability density function of the location of the face, it then satisfies $\int_0^H \int_0^W p(x,y)dxdy = 1$. Plugging in Eq. (1), the EMO score is defined as Eq. (2).

$$EMO = \int_0^H \int_0^W p(x,y) \max_{a \in A} \frac{|B_f \cap B_a|}{|B_f \cup B_a|} dxdy \qquad (2)$$

In practice, we consider the anchor setting according to the Wider Face dataset. We set $S = \{16, 32, 64, 128, 256, 512\}$ to cover the face scale distribution of the dataset, and set $R = \{1\}$ since face boxes are close to squares. Therefore, there are total 6 anchors for each "+" location. In addition, we assume each face can randomly appear anywhere on the image plane with equal probability. Thus, (x, y) are drawn from uniform distributions, i.e. $x \sim U(0, W)$ and $y \sim U(0, H)$.

Since anchors are repeatedly tiled on an image, the overlapping pattern is also periodic w.r.t. the face location. We therefore consider only one period where the face center is enclosed by 4 anchor centers as shown in Figure 3. Then, the face will have the highest IoU with the anchor centered on the closest location to the face center. Due to symmetry, we focus on the cases where the face gets matched to the top-left anchor (dashed red box) with the highest IoU and the blue square shows the $\frac{s_A}{2} \times \frac{s_A}{2}$ region where the face center can appear. Face with the center outside that region will be matched to one of the other three anchors. The relative location of the face center to the anchor center is denoted as (x', y'), where x', y' are drawn from the distribution $U(0, s_A/2)$.

Given a $l \times l$ face with the same size as the anchor box, i.e. $l \in S$, it will be matched to the $l \times l$ anchor. So the IoU score between the face and the matched anchor is

$$IoU = \frac{(l - x')(l - y')}{2l^2 - (l - x')(l - y')} \qquad (3)$$

IoU is a function of the relative location (x', y'). Closer distance from face center to anchor center leads to higher IoU. The EMO score of this face is the expected IoU w.r.t. the distribution of (x', y') derived as in Eq. (4).

$$EMO = \int_0^{\frac{s_A}{2}} \int_0^{\frac{s_A}{2}} (\frac{2}{s_A})^2 \frac{(l - x')(l - y')}{2l^2 - (l - x')(l - y')} dx'dy' \qquad (4)$$

Figure 4 shows the EMO scores given different face scales and anchor strides. It explains why larger faces tend to have higher IoU overlap with anchors. When the face size is fixed, s_A plays an important role in reaching high EMO scores. Given a face, the smaller s_A is, the higher EMO score achieves, especially for small faces. Hence the average max IoU of all faces can be statistically increased.

4. Strategies of New Anchor Design

This section introduces our newly designed anchors, with the purpose of finding more tiny faces. We aim at improving the average IoU especially for tiny faces from the view of theoretically improving EMO score, since average IoU scores are correlated with face recall rate. Based on the analysis in Section 3, we propose to increase the average IoU by reducing anchor stride as well as reducing the distance between the face center and the anchor center.

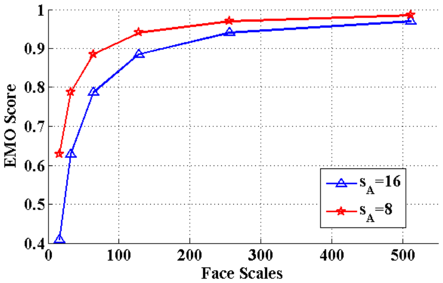


Figure 4. The effect of face scale and anchor stride on the EMO score. Small anchor stride is crucial for tiny faces.

For anchor stride reduction, we look into new network architectures to change the stride of feature map associated with anchors. Three architectures are proposed and discussed in Section 4.1. Additionally, we redefined the anchor locations such that the anchor stride can be further reduced in Section 4.2. Moreover, we propose the face shift jittering method Section 4.3 which can statistically reduce the distance between the face center and the anchor center, the other important factor to increase the IoU overlap.

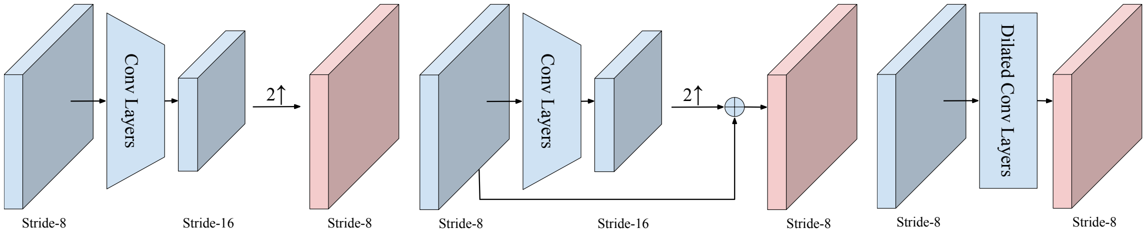
With the aforementioned methods, the EMO score can be improved which theoretically guarantee higher average IoU. However, some very tiny faces are still getting low IoU overlaps. We propose a compensation strategy in Section 4.4 for training which matches very tiny faces to multiple anchors.

4.1. Stride Reduction with Enlarged Feature Maps

As discussed in Section 3.1, anchor stride equals to feature stride in current anchor-based detectors. Therefore, one way to increase the EMO scores is to reduce the anchor stride by enlarging the feature map. This section presents three different architectures that double the height and width of the feature maps as illustrated in Figure 5.

Bilinear upsampling upscales the feature map by a factor of 2 as shown in Figure 5(a). In this network, a deconvolution layer is attached to the feature map and its filters are initialized to have weights of a bilinear upsampler. During training, the filters are updated to adapt to the data.

Figure 5(b) shows the upscaled feature map augmented with the features from shallower large feature map by skip connection. The intuition in this design is to combine high level features for semantical information and low level features for localization precision [18]. In the actual networks, the low-level and high-level feature maps have different numbers of channels. Thus, two 1×1 convolution layers are first added to reduce the number of channels to the same



(a) Bilinear upsampling (b) Bilinear upsampling with skip connection (c) Dilated convolution

Figure 5. Three types of network architecture for reducing the anchor stride by enlarging the feature map (red).

size. Then, after the element-wise addition, another 3 × 3 convolution layer is appended to the final feature map for detection (not shown in Figure 5(b)).

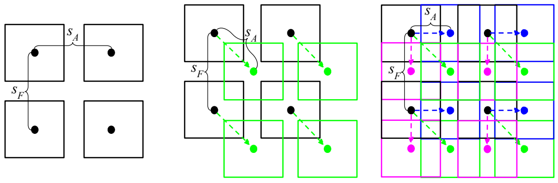
The architectures in Figures 5(a) and 5(b) introduce additional parameters to the networks when enlarging the feature maps, hence increasing the model complexity. However, the same goal can be achieved without additional parameters by using dilated convolutions [23] as shown in Figure 5(c). Specifically, we take out the stride-2 operation (either pooling or convolution) right after the shallower large map and dilate the filters in all the following convolution layers. Note that 1 × 1 filters are not required to be dilated. In addition to not having any additional parameters, dilated convolution also preserve the size of receptive fields of the filters.

With halved anchor stride, the average IoU of tiny faces increases by a large margin as shown in Figure 7, compared to the original one. In addition, we show in Section 5.1 the performance comparison of the three architectures.

4.2. Extra Shifted Anchors

Reducing anchor strides by enlarging feature maps doesn't change the condition that $s_A = s_F$. This section presents a new approach such that $s_A < s_F$. We further reduce s_A by adding extra supportive anchors not centered on the sliding window locations, i.e. *shifted anchors*. This strategy can help to increase the EMO scores without changing the resolution of feature maps. These shifted anchors share the same feature representation with the anchors in the centers.

Specifically, given a feature map with stride s_F, the distance between two adjacent sliding window locations is s_F, and labeled by black dots in Figure 6. In Figure 6(a), each location has a single anchor (black) centered on it, giving the anchor stride of $s_A = s_F$. When extra supportive (green) anchors are added to the bottom-right of the center for all locations, the anchor stride can be reduced to $s_A = s_F/\sqrt{2}$ (Figure 6(b)). In addition, two other supportive anchors (blue and magenta) can be sequentially added to further reduce the anchor stride to $s_A = s_F/2$ (Figure 6(c)). Note that no matter how many anchors are added, all anchors are regularly tiled in the image plane. Indeed, we only need to add small shifted anchors since large anchors already guarantee high average IoUs, which saves the computational time. For example, three shifted anchors of the smallest size (16 × 16) are added on top of enlarged feature maps (Section 4.1) to show further improvement of the average IoUs in Figure 7.



(a) $s_A = s_F$ (b) $s_A = s_F/\sqrt{2}$ (c) $s_A = s_F/2$

Figure 6. Anchor stride reduction by adding shifted anchors.

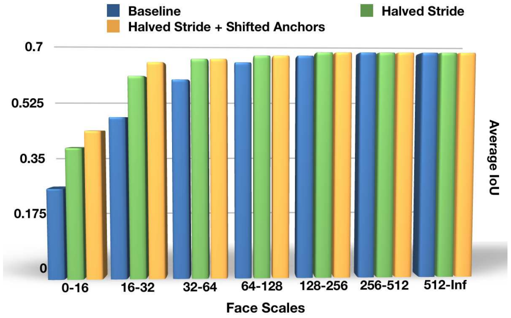


Figure 7. Comparison of the average IoU between original and our anchor design. With our proposed stride reduction techniques, the average IoU is significantly improved for small faces.

4.3. Face Shift Jittering

When computing the EMO score for each face, the center of the face is assumed to be drawn from a 2D uniform distribution. However in the actual datasets, each face has a fixed location. Some of them are closed to the anchor centers so they are more likely to have high IoU overlaps with anchors. While some others far from anchor centers will always get low IoU scores. In order to increase the probability for those faces to get high IoU overlap with anchors, they are randomly shifted in each iteration during training.

Specifically, the image is shifted by a random offset (δ_x, δ_y) in every iteration. δ_x and δ_y are the pixels where the image is shifted to right and down respectively, so that locations of all faces in that image are added by (δ_x, δ_y). The offset is sampled from a discrete uniform distribution, i.e. $\delta_x, \delta_y \in \{0, 1, ..., s_A/2 - 1\}$. We use discrete uniform distribution of offsets to approach the continuous uniform distribution of face locations. We set the maximum offset to be $s_A/2 - 1$ because the period of overlap pattern is $s_A/2$.

4.4. Hard Face Compensation

As shown in Figure 7, even with halved feature strides and shifted small anchors, very tiny faces still have lower average IoU than bigger faces. It is because face scales and locations are continuous whereas anchor scales and locations are discrete. Therefore, there are still some faces whose scales or locations are far away from the anchor. These faces are hard to be matched to anchors.

We propose a compensation strategy of anchor matching to assign hard faces to multiple anchors. Specifically, we first label anchors as positive if their overlapping scores with faces are higher than a threshold T_h, same as the current anchor matching strategy. Then faces whose highest overlapping scores are below T_h are the hard faces. For hard faces the top N anchors with highest overlap with them are labeled as positive. We find the optimal N empirically as described in Section 5.1.

5. 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. Finally the computational time is presented.

5.1. Ablation Study

The Wider Face dataset [43] is used in this ablation study. This dataset has 32,203 images with 393,703 labeled faces with a high degree of variability in scales, occlusions and poses. The images are split into training (40%), validation (10%) and testing (50%) set. Faces in this dataset are classified into Easy, Medium, and Hard subsets according to the difficulties of detection. The hard subset includes a lot of tiny faces. All networks are trained on the training set and evaluated on the validation set. Average Precision (AP) score is used as the evaluation metric. Note that we train and evaluate on the *original* images *without* rescaling since we want to test the detector's ability of finding real tiny faces instead of upscaled tiny faces.

Baseline setup We build an anchor-based detector with ResNet-101 [10] inspired by the R-FCN [4] as our baseline detector. It differs from the original R-FCN in the following aspects. Firstly we set 6 anchors whose scales are from the set $\{16, 32, 64, 128, 256, 512\}$ and all anchors' aspect ratios is 1. This setting matches with the face boxes in the Wider Face dataset. Secondly, "res5" is used to generate region proposals instead of "res4". Thirdly, the threshold of IoU for positive anchors is changed to 0.5. All the other settings follow the original [4]. The baseline detector is trained on the Wider Face training set for 8 epochs. The initial learning rate is set to 0.001 and decreases to 0.0001 after 5 epochs. During training we applied online hard example mining [32] with the ratio between positives and negatives as 1:3. The detector is implemented in the MXNet framework [3] based on the open source code from [5].

The effect of enlarged feature map As discussed in Section 4.1, enlarged feature map reduces the anchor stride so that it helps to increase the EMO scores and the average IoU, especially for tiny faces. To better understand its effect on the final detection performance, we apply three architectures in Figure 5 to the ResNet backbone architecture of the baseline detector. The corresponding stride-8 feature and stride-16 feature in the backbone architecture are "res3" and "res5" respectively. For bilinear upsampling (denoted as **BU**), "res5" is appended with a stride-2 deconvolution layer with filters initialized by a bilinear upsampler. For bilinear upsampling with skip connection (denoted as **BUS**), all the additional convolution layers have 512 output channels. For dilated convolution (denoted as **DC**), all the 3×3 filters in "res4" and "res5" are dilated. Evaluation results are presented as "BU", "BUS" and "DC" in Table 1. Compared to the baseline, the enlarged feature map provides significant improvements on the Hard subset (rising by 7.4% at most), which mainly consists of tiny faces. Among the three architectures, BUS is better than BU at finding tiny faces, but has more false positives, because BUS uses features from the shallower "res3" layer that is less discriminative. DC achieves the best performance on all three subsets without introducing extra parameters. Thus, we fix the architecture to DC in the following experiments.

The effect of additional shifted anchors Adding shifted anchors can further reduce the anchor stride without changing the resolution of the feature map. In Figure 7 we can see with halved anchor stride, faces larger than 32 × 32 pixels already have the same average IoU overlap. So we mainly focus on adding anchors with scales of 16 and 32. As shown in Figure 6, there are two settings of shifted anchors, i.e. added by one to reduce the stride by $1/\sqrt{2}$ or added by three to reduce the stride by $1/2$. We denote the shifted anchor setting as $s \times n$, where $s \in \{16, 32\}$ is the scale of anchor and $n \in \{1, 3\}$ is the number of additional anchors. Noted that we always start with adding anchors of scale 16 since larger anchors cannot affect the average IoU of smaller faces. Hence there are total 5 combinations as presented in the second row section in Table 1. It turns out that adding shifted anchors can improve the AP score on the Hard set. However, there is a trade-off between number of additional anchors and the detection accuracy. More anchors do not always lead to higher average precision, because each anchor is associated with a set of parameters in the network to predict the confidence score and box offsets. As the number of anchors increases, each anchor is matched to fewer faces and the associated parameters are trained from fewer examples. In the end, we find adding just 3 anchors of scale 16 gives the best performance.

The effect of face shift jittering Next we look into the effect of randomly shifting the faces for each iteration during training, denoted as SJ for shift jittering. It is applied to both the model with DC and the model with DC and 3 shifted anchors of scale 16. Experiments show that shift jittering can further improve the AP scores of Hard faces. In Table 1, the AP rises by 0.9% on the model with DC (+DC vs. +DC+SJ) and by 0.3% on the model with DC and 3 shifted anchors of scale 16 (+DC+16x3 vs. +DC+16x3+SJ).

The effect of hard face compensation The hard face compensation completes our final model. It is denoted as HC(N) where N is the number of anchors to which the hard faces are assigned. We find $N = 5$ is a proper choice since smaller N leads to lower recall rate and larger N results in more false positives. To this end, we denote "+DC+16x3+SJ+HC(5)" as our "Final" detector.

The effect of testing size The size of the testing images has significant impact on the face detection precision, especially for tiny faces. Therefore we evaluate our final model on the Hard set of Wider Face validation set with different image sizes, comparing with the state-of-the-art SSH face detector [26]. We show in Table 2 that our detector trained with single scale outperforms the SSH detector trained with multiple scales at every testing size. Note that at the maximum input size of 600x1000, our detector outperforms SSH by 7.1%, showing the effectiveness of our proposed techniques for detecting tiny faces.

The effect of image pyramid Image pyramid for multi-scale training and testing helps improving the detection performance, as shown by "Baseline+Pyramid" in Table 1. By applying our strategies we observe another improvement ("Final+Pyramid"). We follow the same way in [26] to build the pyramid.

5.2. Evaluation on Common Benchmarks

We evaluate our proposed method on the common face detection benchmarks, including Wider Face [43], Annotated Faces in the Wild (AFW) [22], PASCAL Faces [41], and Face Detection Dataset and Benchmark (FDDB) [12]. Our face detector is trained only using Wider Face training set and is tested on those benchmarks *without* further finetuning. We demonstrate consistent state-of-the-art performance across all the datasets. Qualitative results are illustrated in the supplementary materials.

Table 1. Ablative study of each components in our proposed method on Wider Face validation set. Network architectures: **BU** - bilinear upsampling; **BUS** - bilinear upsampling with skip connection; **DC** - dilated convolution. Extra shifted anchors: $s \times n$ - adding n shifted s-by-s anchors. **SJ** - Face Shift Jittering. **HC**(N) - assigning each hard face to top N anchors.

	Easy	Medium	Hard
Baseline	0.934	0.895	0.705
+BU	0.933	0.901	0.710
+BUS	0.926	0.899	0.778
+DC	0.936	0.911	0.779
+DC+16x1	0.934	0.908	0.781
+DC+16x1&32x1	0.936	0.910	0.782
+DC+16x3	0.938	0.912	0.786
+DC+16x3&32x1	0.937	0.909	0.781
+DC+16x3&32x3	0.938	0.913	0.779
+DC+SJ	0.939	0.910	0.788
+DC+16x3+SJ	0.940	0.914	0.789
+DC+16x3+SJ+HC(3)	0.938	0.912	0.793
+DC+16x3+SJ+HC(5) (Final)	**0.940**	**0.914**	**0.795**
+DC+16x3+SJ+HC(7)	0.936	0.912	0.791
Baseline+Pyramid	0.943	0.927	0.840
Final+Pyramid	**0.949**	**0.933**	**0.861**

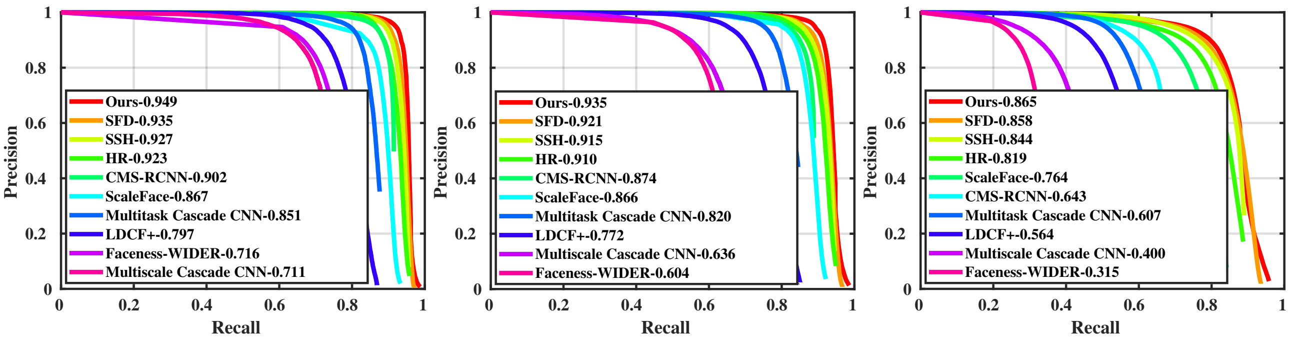
Table 2. The effect of testing size on the average precision (AP) of Hard faces in Wider Face validation set.

Max size	600x1000	800x1200	1200x1600	1400x1800
SSH [26]	0.686	0.784	0.814	0.810
Ours	**0.757**	**0.817**	**0.838**	**0.835**

Wider Face dataset We report the performance of our face detection system on the Wider Face testing set with 16,097 images. Detection results are sent to the database server for receiving the precision-recall curves. Figure 8 illustrates the precision-recall curves along with AP scores. It's clear that our detector consistently achieves best performance on all face cases against recent published face detection methods: SFD [48], SSH [26], ScaleFace [44], HR [11], CMS-RCNN [50], Multitask Cascade CNN [47], LDCF+ [27] and Faceness [42]. The results demonstrate that our proposed components further promote the solution for finding small and hard faces.

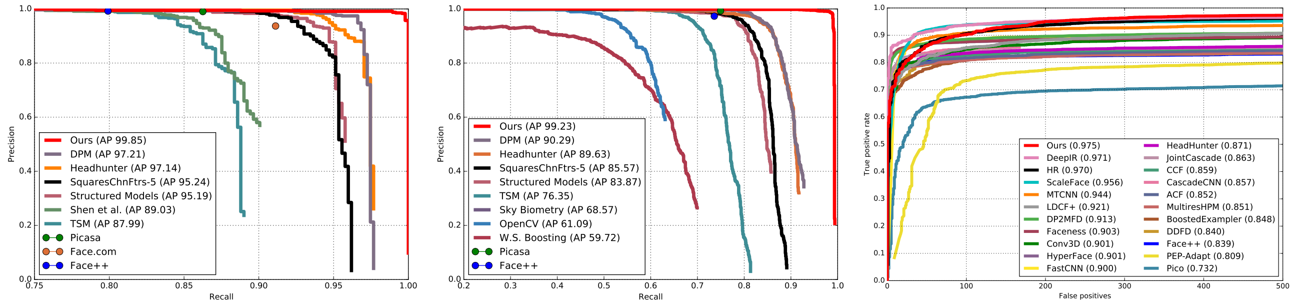
AFW dataset This dataset consists of 205 images with 473 annotated faces. During testing images are kept at the original scale. Our method is compared with popular face detection algorithms [25, 38, 31, 51] as well as commercial face detection systems including Picasa, Face.com and Face++. Figure 9(a) show that our detector outperforms others by a considerable margin.

PASCAL Faces dataset This dataset is a subset of the PASCAL VOC testing set. It has 851 images with 1,335 annotated faces. Our detector is tested on the orig-



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

Figure 8. Precision and recall curves on Wider Face testing set divided into Easy, Medium, and Hard levels.



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

Figure 9. Comparison with popular state-of-the-art methods on the AFW, PASCAL Faces, and FDDB datasets

inal scale of image and compared against popular methods [25, 38, 51] and some commercial face detection systems (Picasa, Face++). The precision-recall curves in Figure 9(b) demonstrate the superiority of our method.

FDDB dataset This dataset has 2,845 images with 5,171 annotated faces. In stead of rectangle boxes, faces in FDDB are represented by ellipses. So we learn a regressor to transform rectangle boxes predicted by our detector to ellipses. Again, we test on the original images without rescaling. We compare our approach with other state-of-the-art methods [34, 11, 44, 47, 27, 28, 42, 17, 29, 35, 25, 2, 40, 15, 39, 9, 14, 7, 49, 13, 24] which *don't* add additional self-labeled annotations. As shown in Figure 9(c), our detector achieves high recall rate and best AP score.

5.3. Runtime Speed

This section reports the runtime speed of our detector on the Wider Face validation set. We vary the image size to get a set of AP scores with corresponded average inference times as shown in Figure 10. Our detector forms an upper envelope of other detectors [20, 30, 44, 11, 26]. Some numbers are acquired from [44]. All detectors are evaluated using a single NVIDIA Titan X GPU with batch size 1.

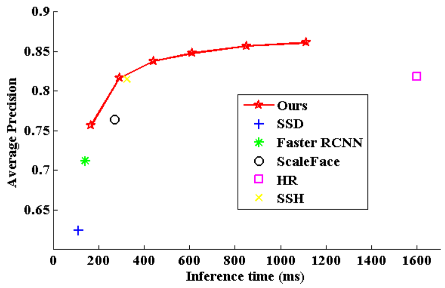


Figure 10. Speed (ms) vs. accuracy (AP) on Wider Face Hard set.

6. Conclusion

This work identified low face-anchor overlap as the major reason hindering anchor-based detectors to detect tiny faces. We proposed the new EMO score to characterize anchors' capability of getting high overlaps with faces, providing an in-depth analysis of the anchor matching mechanism. This inspired us to come up with several simple but effective strategies of a new anchor design for higher face IoU scores. Consequently, our method outperforms the baseline anchor-based detector by a considerable margin and achieves the state-of-the-art performance on the challenging Wider Face, AFW, PASCAL Faces and FDDB datasets.

References

[1] C. Bhagavatula, C. Zhu, K. Luu, and M. Savvides. Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[2] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun. Joint cascade face detection and alignment. In *Computer Vision–ECCV 2014*, pages 109–122. Springer, 2014. 2, 8

[3] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *NIPS Workshop on Machine Learning Systems (LearningSys)*, 2016. 6

[4] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems*, pages 379–387, 2016. 1, 2, 6

[5] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. *International Conference on Computer Vision (ICCV)*, 2017. 6

[6] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. volume 88, pages 303–338. Springer, 2010. 1

[7] S. S. Farfade, M. J. Saberian, and L.-J. Li. Multi-view face detection using deep convolutional neural networks. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 643–650. ACM, 2015. 8

[8] P. Felzenszwalb, R. Girshick, and D. McAllester. Cascade object detection with deformable part models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 1, 2

[9] G. Ghiasi and C. C. Fowlkes. Occlusion coherence: Detecting and localizing occluded faces. *arXiv preprint arXiv:1506.08347*, 2015. 2, 8

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2, 6

[11] P. Hu and D. Ramanan. Finding tiny faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 1, 2, 7, 8

[12] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010. 2, 7

[13] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic part model for unsupervised face detector adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 793–800, 2013. 8

[14] H. Li, Z. Lin, J. Brandt, X. Shen, and G. Hua. Efficient boosted exemplar-based face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1843–1850, 2014. 8

[15] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5325–5334, 2015. 8

[16] J. Li and Y. Zhang. Learning surf cascade for fast and accurate object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3468–3475, 2013. 1, 2

[17] Y. Li, B. Sun, T. Wu, and Y. Wang. face detection with end-to-end integration of a convnet and a 3d model. In *European Conference on Computer Vision*, pages 420–436. Springer, 2016. 1, 2, 8

[18] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 4

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *ECCV*, 2016. 1, 2, 8

[21] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 2017. 1

[22] P. M. R. M. Koestinger, P. Wohlhart and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. 2011. 2, 7

[23] S. Mallat. *A wavelet tour of signal processing*. Academic press, 1999. 5

[24] N. Markuš, M. Frljak, I. S. Pandzic, J. Ahlberg, and R. Forchheimer. A method for object detection based on pixel intensity comparisons. In *2nd Croatian Computer Vision Workshop*, 2013. 8

[25] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *Computer Vision–ECCV 2014*, pages 720–735. Springer, 2014. 2, 7, 8

[26] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis. Ssh: Single stage headless face detector. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 2, 7, 8

[27] E. Ohn-Bar and M. M. Trivedi. To boost or not to boost? on the limits of boosted trees for object detection. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 3350–3355. IEEE, 2016. 1, 2, 7, 8

[28] R. Ranjan, V. M. Patel, and R. Chellappa. A deep pyramid deformable part model for face detection. In *Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference on*, pages 1–8. IEEE, 2015. 8

[29] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *arXiv preprint arXiv:1603.01249*, 2016. 8

[30] S. Ren, K. He, R. Girshick, and J. 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, 8

[31] X. Shen, Z. Lin, J. Brandt, and Y. Wu. Detecting and aligning faces by image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3460–3467, 2013. 7

[32] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769, 2016. 6

[33] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 1, 2

[34] X. Sun, P. Wu, and S. C. Hoi. Face detection using deep learning: An improved faster rcnn approach. *arXiv preprint arXiv:1701.08289*, 2017. 2, 8

[35] D. Triantafyllidou and A. Tefas. A fast deep convolutional neural network for face detection in big visual data. In *INNS Conference on Big Data*, pages 61–70. Springer, 2016. 8

[36] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2001. 1, 2

[37] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016. 1

[38] J. Yan, X. Zhang, Z. Lei, and S. Z. Li. Face detection by structural models. *Image and Vision Computing*, 32(10):790–799, 2014. 7, 8

[39] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Aggregate channel features for multi-view face detection. In *Biometrics (IJCB), 2014 IEEE International Joint Conference on*, pages 1–8. IEEE, 2014. 8

[40] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Convolutional channel features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 82–90, 2015. 8

[41] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Fine-grained evaluation on face detection in the wild. In *Automatic Face and Gesture Recognition (FG), 11th IEEE International Conference on*. IEEE, 2015. 2, 7

[42] S. Yang, P. Luo, C.-C. Loy, and X. Tang. From facial parts responses to face detection: A deep learning approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3676–3684, 2015. 1, 2, 7, 8

[43] S. Yang, P. Luo, C.-C. Loy, and X. Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5525–5533, 2016. 2, 6, 7

[44] S. Yang, Y. Xiong, C. C. Loy, and X. Tang. Face detection through scale-friendly deep convolutional networks. *arXiv preprint arXiv:1706.02863*, 2017. 1, 2, 7, 8

[45] X. Yu, J. Huang, S. Zhang, W. Yan, and D. Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1944–1951, 2013. 1, 2

[46] C. Zhang and Z. Zhang. A survey of recent advances in face detection. Technical Report MSR-TR-2010-66, June 2010. 2

[47] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 1, 2, 7, 8

[48] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. S3fd: Single shot scale-invariant face detector. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 2, 7

[49] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 386–391, 2013. 8

[50] C. Zhu, Y. Zheng, K. Luu, and M. Savvides. Cms-rcnn: contextual multi-scale region-based cnn for unconstrained face detection. In *Deep Learning for Biometrics*, pages 57–79. Springer, 2017. 1, 2, 7

[51] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886, 2012. 1, 2, 7, 8