

# DSFD: Dual Shot Face Detector

Jian Li<sup>†\*</sup> Yabiao Wang<sup>‡</sup> Changan Wang<sup>‡</sup> Ying Tai<sup>‡</sup>  
Jianjun Qian<sup>†</sup> Jian Yang<sup>†</sup> Chengjie Wang<sup>‡</sup> Jilin Li<sup>‡</sup> Feiyue Huang<sup>‡</sup>  
<sup>†</sup>Nanjing University of Science and Technology      <sup>‡</sup>YouTu Lab, Tencent

<sup>†</sup>lijiannuist@gmail.com, {csjqian, csjyang}@njust.edu.cn

<sup>‡</sup>{casewang, changanwang, yingtai, jasoncjwang, jerolinli, garyhuang}@tencent.com



Figure 1: **Visual results of our DSFD.** Our method is robust to various variations on scale, blurry, illumination, pose, occlusion, reflection, makeup, etc.

## Abstract

Recently, Convolutional Neural Network (CNN) has achieved great success in face detection. However, it remains a challenging problem for the current face detection methods owing to high degree of variability in scale, pose, occlusion, expression, appearance and illumination. In this paper, we propose a novel face detection network named Dual Shot Face Detector(DSFD), which inherits the architecture of SSD and introduces a Feature Enhance Module (FEM) for transferring the original feature maps to extend the single shot detector to dual shot detector. Specially, Progressive Anchor Loss (PAL) computed by using two set of anchors is adopted to effectively facilitate the features. Additionally, we propose an Improved Anchor Matching (IAM) method by integrating novel data augmentation techniques

and anchor design strategy in our DSFD to provide better initialization for the regressor. Extensive experiments on popular benchmarks: WIDER FACE (easy: 0.966, medium: 0.957, hard: 0.904) and FDDB (discontinuous: 0.991, continuous: 0.862) demonstrate the superiority of DSFD over the state-of-the-art face detectors (e.g., PyramidBox and SRN). Code will be made available upon publication.

## 1. Introduction

Face detection is a fundamental step for various facial applications, like face alignment, parsing, recognition, and verification. As the pioneering work for face detection, Viola-Jones [23] adopts AdaBoost algorithm with hand-crafted features, which are now replaced by deeply learned features from the convolutional neural network (CNN) [8] that achieves great progress. Although the CNN based face

\*This work was done when Jian Li was an intern at Tencent YouTu Lab.

detectors have been extensively studied, detecting faces with high degree of variability in scale, pose, occlusion, expression, appearance and illumination in real-world scenarios remains a challenge.

Previous state-of-the-art face detectors can be roughly divided into two categories. The first one is mainly based on the Region Proposal Network (RPN) adopted in Faster R-CNN [19] and employs two stage detection schemes [24, 27, 29]. RPN is trained end-to-end and generates high-quality region proposals which are further refined by Fast R-CNN detector. The other one is Single Shot Detector (SSD) [15] based one-stage methods, which get rid of RPN, and directly predict the bounding boxes and confidence [3, 21, 32]. Recently, one-stage face detection framework has attracted more attention due to its higher inference efficiency and straightforward system deployment.

Despite the progress achieved by the above face detectors, there are still some problems existed in the following three aspects:

**Feature learning** Feature extraction part is essential for a face detector. Currently, Feature Pyramid Network (FPN) [12] is widely used in state-of-the-art face detectors for rich features. However, FPN just aggregates hierarchical feature maps between high and low-level output layers, which does not consider the current layers information, and the context relationship between anchors is ignored.

**Loss design** The conventional loss functions used in object detection include a regression loss for the face region and a classification loss for identifying if a face is detected or not. To further address the class imbalance problem, Lin et al. [13] proposed Focal Loss to focus training on a sparse set of hard examples. To use all original and enhanced features, Zhang et al. proposed Hierarchical Loss to effectively learn the network [30]. However, the above loss functions do not consider progressive learning ability of feature maps in different levels.

**Anchor matching** Basically, pre-set anchors for each feature map are generated by regularly tiling a collection of boxes with different scales and aspect ratios on the image. Some works [21, 32] analyze a series of reasonable anchor scales and anchor compensation strategy to increase positive anchors' number. However, such strategy ignores random sampling in data augmentation. Continuous face scale and a large number of discrete anchor scales still make huge ratio differences of negative and positive anchors.

To address the above three issues, we propose a novel network based on the SSD pipeline named Dual Shot Face Detection (DSFD). First, combining the similar setting of low-level FPN in PyramidBox and the Receptive Field Block (RFB) in RFBNet [14], we introduce a Feature Enhance Module (FEM) to enhance the discriminability and robustness of the features. Second, motivated by the hierarchical loss [30] and pyramid anchor [21] in PyramidBox,

we propose Progressive Anchor Loss (PAL) that computes auxiliary supervision loss by a set of smaller anchors to effectively facilitate the original features, since smaller anchor tiled to original feature maps cell may have more semantic information for classification and high-resolution location information for small faces. Last but not least, we propose an Improved Anchor Matching (IAM) method, which integrates anchor partition strategy and anchor-based data augmentation techniques into our DSFD to match anchors and ground truth faces as far as possible to provide better initialization for the regressor. Fig. 1 shows the effectiveness of our proposed DSFD on various variations, especially on extreme small faces or heavily occluded faces.

In summary, the main contributions of this paper include:

- A novel Feature Enhance Module to utilize different level information and thus obtain more discriminability and robustness features.
- Auxiliary supervisions introduced in early layers by using a set of smaller anchors to effectively facilitate the features.
- An improved anchor matching strategy to match anchors and ground truth faces as far as possible to provide better initialization for the regressor.
- Comprehensive experiments conducted on popular benchmarks FDDB and WIDER FACE to demonstrate the superiority of our proposed DSFD network compared with the state-of-the-art methods.

## 2. Related work

We review the prior works from three perspectives.

**Feature Learning** Early works on face detection mainly rely on hand-crafted features, such as Harr-like features [23], control point set [1], edge orientation histograms [10]. However, hand-crafted features design is lack of guidance. With the great progress of deep learning, hand-crafted features have been replaced by Convolutional Neural Networks (CNN). For example, Overfeat [20], Cascade-CNN [11], MTCNN [31] adopt CNN as a sliding window detector on image pyramid to build feature pyramid. However, using an image pyramid is slow and memory inefficient. As the result, most two stage detectors extract features on single scale. R-CNN [5, 6] obtains region proposals by selective search [22], and then forwards each normalized image region through a CNN to classify. Faster R-CNN [19], R-FCN [4] employ Region Proposal Network (RPN) to generate initial region proposals. Besides, ROI-pooling [19] and position-sensitive ROI pooling [4] are applied to extract features from each region.

More recently, some research indicates that multi-scale features perform better for tiny objects. Specifically, SSD [15], MS-CNN [2], SSH [18], S3FD [32] predict boxes on multiple layers of feature hierarchy. FCN [17], Hypercolumns [7], Parsenet [16] fuse multiple layer features in

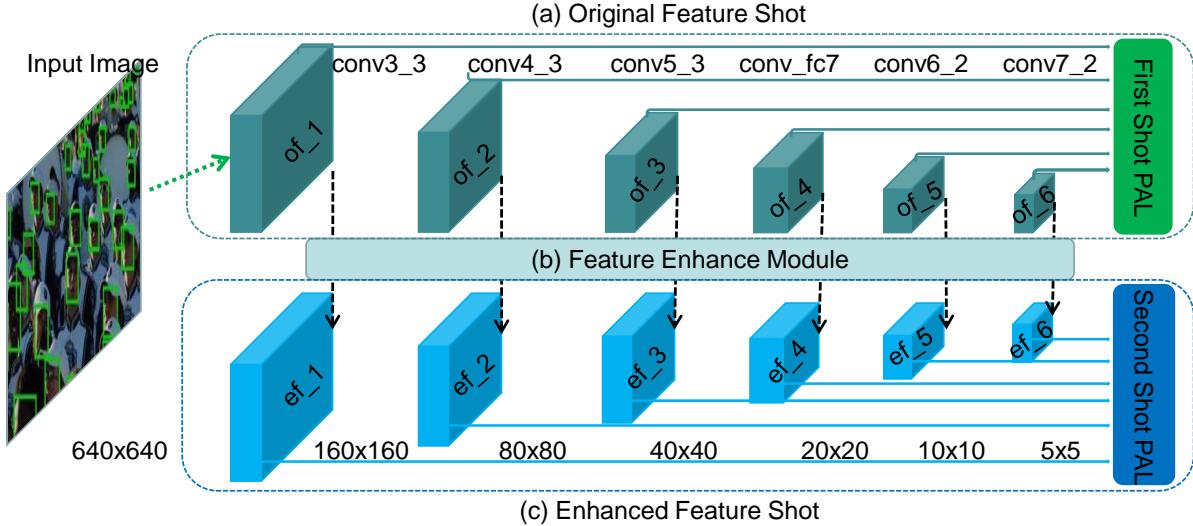


Figure 2: Our DSFD framework uses a Feature Enhance Module (b) on top of a feedforward VGG16 architecture to generate the enhanced features (c) from the original features (a), along with two loss layers named first shot PAL for the original features and second shot PAL for the enhanced features.

segmentation. FPN [12], a top-down architecture, integrate high-level semantic information to all scales. FPN-based methods, such as FAN [25], PyramidBox [21] achieve significant improvement on detection. However, these methods do not consider the current layers information. Different from the above methods that ignore the context relationship between anchors, we propose a feature enhance module that incorporates multi-level dilated convolutional layers to enhance the semantic of the features.

**Loss Design** Generally, the objective loss in detection is a weighted sum of classification loss (*e.g.* softmax loss) and box regression loss (*e.g.*  $L_2$  loss). Girshick et al. [5] propose smooth  $L_1$  loss to prevent exploding gradients. Lin et al. [13] discover that the class imbalance is one obstacle for better performance in one stage detector, hence they propose focal loss, a dynamically scaled cross entropy loss. Besides, Wang et al. [26] design RepLoss for pedestrian detection, which improves performance in occlusion scenarios. FANet [30] create a hierarchical feature pyramid and presents hierarchical loss for their architecture. However, the anchors used in FANet are kept the same size in different stages. In this work, we adaptively choose different anchor sizes in different stages to facilitate the features.

**Anchor Matching** To make the model more robust, most detection methods [15, 28, 32] do data augmentation, such as color distortion, horizontal flipping, random crop and multi-scale training. Zhang et al. [32] propose an anchor compensation strategy to make tiny faces to match enough anchors during training. Wang et al. [28] propose random crop to generate large number of occluded faces for training. However, these methods ignore random sampling in data augmentation, while ours combines anchor assign to provide better data initialization for anchor matching.

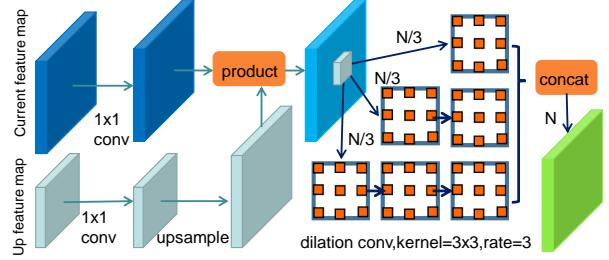


Figure 3: Feature Enhance Module illustrating the current feature map cell interactive with neighbor in current feature maps and up feature maps.

### 3. Dual Shot Face Detector

We firstly introduce the pipeline of our proposed framework DSFD, and then detailly describe our feature enhance module in Sec. 3.2, progressive anchor loss in Sec. 3.3 and improved anchor matching in Sec. 3.4, respectively.

#### 3.1. Pipeline of DSFD

The framework of DSFD is illustrated in Fig. 2. Our architecture uses the same extended VGG16 backbone as PyramidBox [21] and S3FD [32], which is truncated before the classification layers and added with some auxiliary structures. We select conv3\_3, conv4\_3, conv5\_3, conv\_fc7, conv6\_2 and conv7\_2 as the first shot detection layers to generate six original feature maps named  $of_1, of_2, of_3, of_4, of_5, of_6$ . Then, our proposed FEM transfers these original feature maps into six enhanced feature maps named  $ef_1, ef_2, ef_3, ef_4, ef_5, ef_6$ , which have the same sizes as the original ones and are fed into SSD-style head to construct the second shot detection layers. Note that the input size of the training image is 640, which means the

feature map size of the lowest-level layer to highest-level layer is from 160 to 5. Different from S3FD and Pyramid-Box, after we utilize the receptive field enlargement in FEM and the new anchor design strategy, its unnecessary for the three sizes of stride, anchor and receptive field to satisfy equal-proportion interval principle. Therefore, our DSFD is more flexible and robustness. Besides, the original and enhanced shots have two different losses, respectively named First Shot progressive anchor Loss (FSL) and Second Shot progressive anchor Loss (SSL).

### 3.2. Feature Enhance Module

Feature Enhance Module is able to enhance original features to make them more discriminable and robust, which is called FEM for short. For enhancing original neuron cell  $oc_{(i,j,l)}$ , FEM utilizes different dimension information including upper layer original neuron cell  $oc_{(i,j,l)}$  and current layer non-local neuron cells:  $nc_{(i-\varepsilon,j-\varepsilon,l)}$ ,  $nc_{(i-\varepsilon,j,l)}$ , ...,  $nc_{(i,j+\varepsilon,l)}$ ,  $nc_{(i+\varepsilon,j+\varepsilon,l)}$ . Specially, the enhanced neuron cell  $ec_{(i,j,l)}$  can be mathematically defined as follow:

$$\begin{aligned} ec_{(i,j,l)} &= f_{concat}(f_{dilation}(nc_{(i,j,l)})) \\ nc_{i,j,l} &= f_{prod}(oc_{(i,j,l)}, f_{up}(oc_{(i,j,l+1)})) \end{aligned} \quad (1)$$

where  $c_{i,j,l}$  is a cell located in  $(i, j)$  coordinate of the feature maps in the  $l$ -th layer,  $f$  denotes a set of basic dilation convolution, elem-wise production, up-sampling or concatenation operations. Fig. 3 illustrates the idea of FEM, which is inspired by FPN [12] and RFB [14]. Here, we first use  $1 \times 1$  convolutional kernel to normalize the feature maps. Then, we up-sample upper feature maps to do element-wise product with the current ones. Finally, we split the feature maps to three parts, followed by three sub-networks containing different numbers of dilation convolutional layers.

### 3.3. Progressive Anchor Loss

In this subsection, we adopt the multi-task loss [15, 19] since it helps to facilitate the original and enhanced feature maps training task in two shots. First, our Second Shot anchor-based multi-task Loss function is defined as:

$$\begin{aligned} \mathcal{L}_{SSL}(p_i, p_i^*, t_i, g_i, a_i) &= \frac{1}{N_{conf}} (\Sigma_i L_{conf}(p_i, p_i^*)) \\ &\quad + \frac{\beta}{N_{loc}} \Sigma_i p_i^* L_{loc}(t_i, g_i, a_i), \end{aligned} \quad (2)$$

where  $N_{conf}$  and  $N_{loc}$  indicate the number of positive and negative anchors, and the number of positive anchors respectively,  $L_{conf}$  is the softmax loss over two classes (face vs. background), and  $L_{loc}$  is the smooth  $L_1$  loss between the parameterizations of the predicted box  $t_i$  and ground-truth box  $g_i$  using the anchor  $a_i$ . When  $p_i^* = 1$  ( $p_i^* = \{0, 1\}$ ), the anchor  $a_i$  is positive and the localization loss is activated.  $\beta$  is a weight to balance the effects of the two terms.

Table 1: The stride size, feature map size, anchor scale, anchor ratio, anchor number of six original and enhanced features for two shots.

Feature	Stride	Size	Scale	Ratio	Number
ef.1 (of_1)	4	$160 \times 160$	16 (8)	1.5 : 1	25600
ef.2 (of_2)	8	$80 \times 80$	32 (16)	1.5 : 1	6400
ef.3 (of_3)	16	$40 \times 40$	64 (32)	1.5 : 1	1600
ef.4 (of_4)	32	$20 \times 20$	128 (64)	1.5 : 1	400
ef.5 (of_5)	64	$10 \times 10$	256 (128)	1.5 : 1	100
ef.6 (of_6)	128	$5 \times 5$	512 (256)	1.5 : 1	25

Compared to the enhanced feature maps in the same level, the original feature maps have less semantic information for classification but more high resolution location information for detection. Therefore, we believe that the original feature maps can detect and classify smaller faces. As the result, we propose the First Shot multi-task Loss with a set of smaller anchors as follows:

$$\begin{aligned} \mathcal{L}_{FSL}(p_i, p_i^*, t_i, g_i, sa_i) &= \frac{1}{N_{conf}} \Sigma_i L_{conf}(p_i, p_i^*) \\ &\quad + \frac{\beta}{N_{loc}} \Sigma_i p_i^* L_{loc}(t_i, g_i, sa_i), \end{aligned} \quad (3)$$

where  $sa$  indicates the smaller anchors in the first shot layers, and the two shots losses can be weighted summed into a whole Progressive Anchor Loss as follows:

$$\mathcal{L}_{PAL} = \mathcal{L}_{FSL}(sa) + \lambda \mathcal{L}_{SSL}(a). \quad (4)$$

Note that anchor size in the first shot is half of ones in the second shot, and  $\lambda$  is weight factor. Detailed assignment on the anchor size is described in Sec. 3.4. In prediction process, we only use the output of the second shot, which means no additional computational cost is introduced.

### 3.4. Improved Anchor Matching

During training, we need to compute positive and negative anchors and determine which anchor corresponds to its face bounding box. Current anchor matching method is bidirectional between the anchor and ground-truth face. Therefore, anchor design and face sampling during augmentation are collaborative to match the anchors and faces as far as possible for better initialization of the regressor.

Table 1 shows details of our anchor design on how each feature map cell is associated to the fixed shape anchor. We set anchor ratio 1.5:1 based on face scale statistics. Anchor size for the original feature is one half of the enhanced feature. Additionally, with probability of 2/5, we utilize anchor-based sampling like data-anchor-sampling in PyramidBox, which randomly selects a face in an image, crops sub-image containing the face, and sets the size ratio between sub-image and selected face to 640/rand (16, 32, 64, 128, 256, 512). For the remaining 3/5 probability, we adopt data augmentation similar to SSD [15]. In

Table 2: Effectiveness of Feature Enhance Module on the AP performance.

Component	Easy	Medium	Hard
FSSD+VGG16	92.6%	90.2%	79.1%
FSSD+VGG16+FEM	<b>93.0%</b>	<b>91.4%</b>	<b>84.6%</b>

Table 3: Effectiveness of Progressive Anchor Loss on the AP performance.

Component	Easy	Medium	Hard
FSSD+RES50	93.7%	92.2%	81.8%
FSSD+RES50+FEM	95.0%	94.1%	88.0%
FSSD+RES50+FEM+PAL	<b>95.3%</b>	<b>94.4%</b>	<b>88.6%</b>

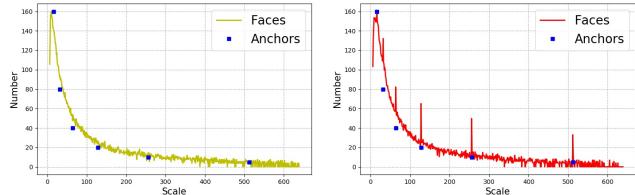


Figure 4: The number distribution of different scales of faces compared between traditional anchor matching (left) and our improved anchor matching (right).

order to improve the recall rate of faces and ensure anchor classification ability simultaneously, we set Intersection-over-Union (IoU) threshold 0.4 to assign anchor to its ground-truth faces.

## 4. Experiments

### 4.1. Implementation Details

First, we present the details in implementing our network. The backbone networks are initialized by the pre-trained VGG/ResNet on ImageNet. All newly added convolution layers' parameters are initialized by the ‘xavier’ method. We use SGD with 0.9 momentum, 0.0005 weight decay to fine-tune our DSFD model. The batch size is set to 16. The learning rate is set to  $10^{-3}$  for the first 40k steps, and we decay it to  $10^{-4}$  and  $10^{-5}$  for two 10k steps.

During inference, the first shot's outputs are ignored and the second shot predicts top 5k high confident detections. Non-maximum suppression is applied with jaccard overlap of 0.3 to produce top 750 high confident bounding boxes per image. For 4 bounding box coordinates, we round down top left coordinates and round up width and height to expand the detection bounding box.

### 4.2. Analysis on DSFD

In this subsection, we conduct extensive experiments and ablation studies on the WIDER FACE dataset to evaluate the effectiveness of several contributions of our proposed framework, including feature enhance module, progressive anchor loss, and improved anchor matching. For fair comparisons, we use the same parameter settings for all the ex-

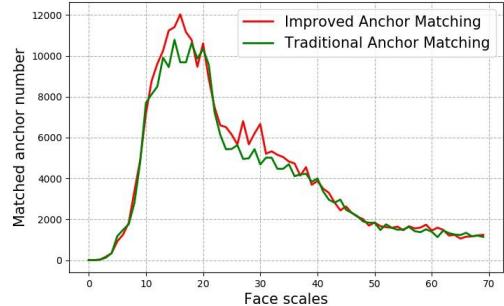


Figure 5: Comparisons on number distribution of matched anchor for ground truth faces between traditional anchor matching (blue line) and our improved anchor matching (red line).

periments, except for the specified changes to the components. All models are trained on the WIDER FACE training set and evaluated on validation set. To better understand DSFD, we select different baselines to ablate each component on how this part affects the final performance.

**Feature Enhance Module** First, We adopt anchor designed in S3FD [32], PyramidBox [21] and six original feature maps generated by VGG16 to perform classification and regression, which is named Face SSD (FSSD) as the baseline. We then use VGG16-based FSSD as the baseline to add feature enhance module for comparison. Table 2 shows that our feature enhance module can improve VGG16-based FSSD from 92.6%, 90.2%, 79.1% to 93.0%, 91.4%, 84.6%.

**Progressive Anchor Loss** Second, we use Res50-based FSSD as the baseline to add progressive anchor loss for comparison. We use four residual blocks' outputs in ResNet to replace the outputs of conv3\_3, conv4\_3, conv5\_3, conv\_fc7 in VGG. Except for VGG16, we do not perform layer normalization. Table 3 shows our progressive anchor loss can improve Res50-based FSSD using FEM from 95.0%, 94.1%, 88.0% to 95.3%, 94.4%, 88.6%.

**Improved Anchor Matching** To evaluate our improved anchor matching strategy, we use Res101-based FSSD without anchor compensation as the baseline. Table 4 shows that our improved anchor matching can improve Res101-based FSSD using FEM from 95.8%, 95.1%, 89.7% to 96.1%, 95.2%, 90.0%. Finally, we can improve our DSFD to 96.6%, 95.7%, 90.4% with ResNet152 as the backbone.

Besides, Fig. 4 shows that our improved anchor matching strategy greatly increases the number of ground truth faces that are closed to the anchor, which can reduce the contradiction between the discrete anchor scales and continuous face scales. Moreover, Fig. 5 shows the number distribution of matched anchor number for ground truth faces, which indicates our improved anchor matching can significantly increase the matched anchor number, and the averaged number of matched anchor for different scales of faces can be improved from 6.4 to about 6.9.

From the above analysis and results, some promising conclusions can be drawn: 1) Feature enhance is crucial.

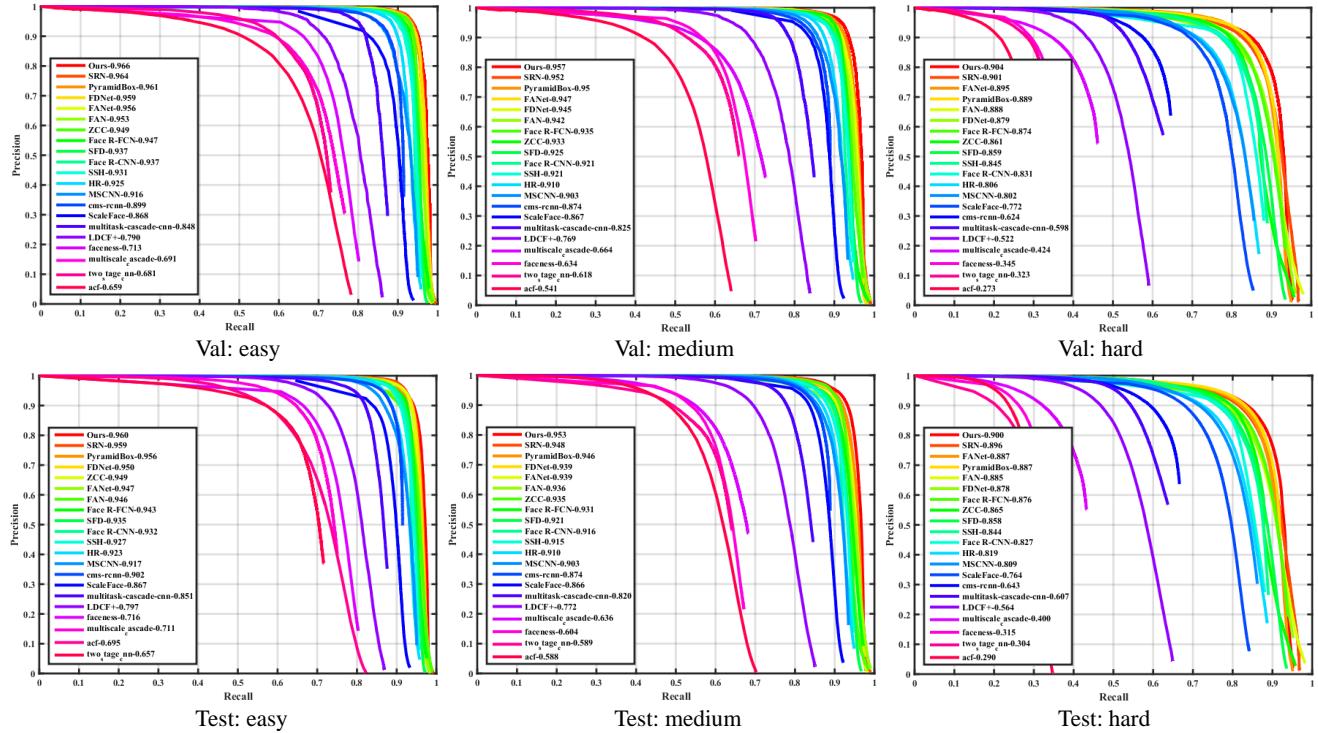


Figure 6: Precision-recall curves on WIDER FACE validation and testing subset.

Table 4: Effectiveness of Improved Anchor Matching on the AP performance.

Component	Easy	Medium	Hard
FSSD+RES101	95.1%	93.6%	83.7%
FSSD+RES101+FEM	95.8%	95.1%	89.7%
FSSD+RES101+FEM+IAM	96.1%	95.2%	90.0%
FSSD+RES101+FEM+IAM+PAL	96.3%	95.4%	90.1%
FSSD+RES152+FEM+IAM+PAL	<b>96.6%</b>	<b>95.7%</b>	90.4%
FSSD+RES152+FEM+IAM+PAL+LargeBS	96.4%	<b>95.7%</b>	<b>91.2%</b>

Table 5: Effectiveness of different backbones.

Component	Params	ACC@Top-1	Easy	Medium	Hard
FSSD+RES101+FEM+IAM+PAL	399M	77.44%	96.3%	95.4%	90.1%
FSSD+RES152+FEM+IAM+PAL	459M	78.42%	<b>96.6%</b>	<b>95.7%</b>	<b>90.4%</b>
FSSD+SE-RES101+FEM+IAM+PAL	418M	78.39%	95.7%	94.7%	88.6%
FSSD+DPN98+FEM+IAM+PAL	515M	79.22%	96.3%	95.5%	<b>90.4%</b>
FSSD+SE-RESNeXt101_32×4d+FEML+IAM+PA	416M	80.19%	95.7%	94.8%	88.9%

We use a more robust and discriminative feature enhance module to improve the feature presentation ability, especially for hard face. 2) Auxiliary loss based on progressive anchor is used to train all 12 different scale detection feature maps, and it improves the performance on easy, medium and hard faces simultaneously. 3) Our improved anchor matching provides better initial anchors and ground-truth faces to regress anchor from faces, which achieves the improvements of 0.3%, 0.1%, 0.3% on three settings, respectively. Additionally, when we enlarge the training batch size (*i.e.*, LargeBS), the result in hard setting can get 91.2% AP.

**Effects of Different Backbones** To better understand our DSFD, we further conducted experiments to examine how different backbones affect classification and detection performance. Specifically, we use the same setting except for the feature extraction network, we implement SE-ResNet101, DPN-98, SE-ResNeXt101\_32×4d following the ResNet101 setting in our DSFD. From Table 5, DSFD with SE-ResNeXt101\_32×4d got 95.7%, 94.8%, 88.9%, on easy, medium and hard settings respectively, which indicates that more complexity model and higher Top-1 ImageNet classification accuracy may not benefit face detection

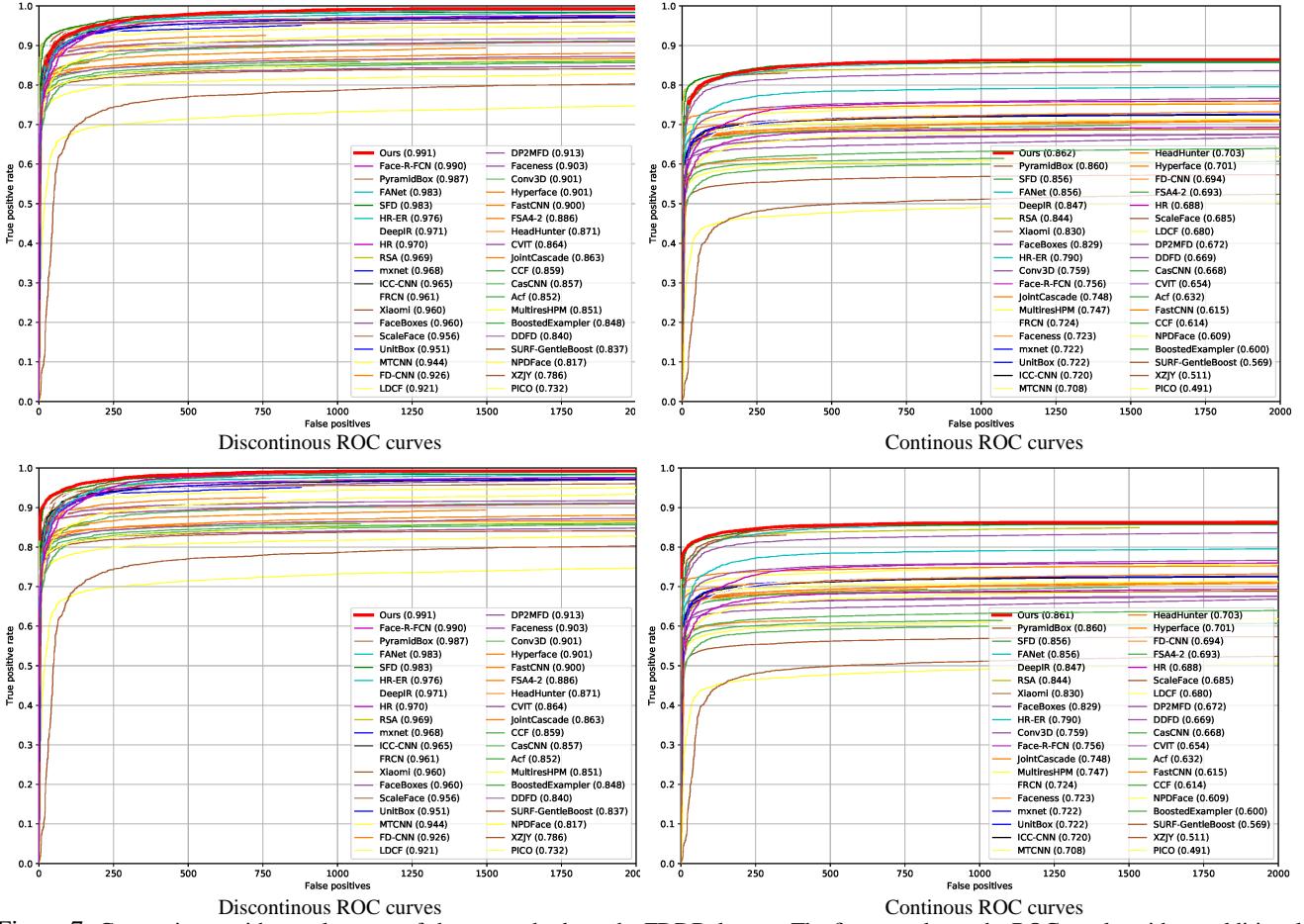


Figure 7: Comparisons with popular state-of-the-art methods on the FDDB dataset. The first row shows the ROC results without additional annotations, and the second row shows the ROC results with additional annotations.

AP. Therefore, in our DSFD framework, better performance on classification are not necessary for better performance on detection. Our DSFD enjoys high inference speed benefited from simply using the second shot detection results. For VGA resolution inputs to Res50-based DSFD, it runs 22 FPS on NVIDIA GPU P40 during inference.

### 4.3. Comparisons with State-of-the-Art Methods

We evaluate the proposed DSFD on two popular face detection benchmarks, including WIDER FACE [28] and Face Detection Data Set and Benchmark (FDDB) [9]. Our model is trained only using the training set of WIDER FACE, and then evaluated on both benchmarks without any further fine-tuning. We also follow the similar way used in [25] to build the image pyramid for multi-scale testing and use more powerful backbone similar as [3].

**WIDER FACE Dataset** It contains 393,703 annotated faces with large variations in scale, pose and occlusion in total 32,203 images. For each of the 60 event classes, 40%, 10%, 50% images of the database are randomly selected as training, validation and testing sets. Besides, each sub-

set is further defined into three levels of difficulty: 'Easy', 'Medium', 'Hard' based on the detection rate of a baseline detector. As shown in Fig. 6, our DSFD achieves the best performance among all of the state-of-the-art face detectors based on the average precision (AP) across the three subsets, *i.e.*, 96.6% (Easy), 95.7% (Medium) and 90.4% (Hard) on validation set, and 96.0% (Easy), 95.3% (Medium) and 90.0% (Hard) on test set. Fig. 8 shows more examples to demonstrate the effects of DSFD on handling faces with various variations, in which the blue bounding boxes indicate the detector confidence is above 0.8.

**FDDB Dataset** It contains 5,171 faces in 2,845 images taken from the faces in the wild data set. Since WIDER FACE has bounding box annotation while faces in FDDB are represented by ellipses, we learn a post-hoc ellipses regressor to transform the final prediction results. As shown in Fig. 7, our DSFD achieves state-of-the-art performance on both discontinuous and continuous ROC curves, *i.e.* 99.1% and 86.2% when the number of false positives equals to 1,000. After adding additional annotations to those unlabeled faces [32], the false positives of our model can be



Figure 8: Effectiveness of our DSFD to various large variations on scale, pose, occlusion, blurry, makeup, illumination, modality and reflection. Blue bounding boxes indicate the detector confidence is above 0.8.

further reduced and outperform all other methods.

## 5. Conclusions

This paper introduces a novel face detector named Dual Shot Face Detector (DSFD). In this work, we propose a novel Feature Enhance Module that utilizes different level information and thus obtains more discriminability and robustness features. Auxiliary supervisions introduced in

early layers by using smaller anchors are adopted to effectively facilitate the features. Moreover, an improved anchor matching method is introduced to match anchors and ground truth faces as far as possible to provide better initialization for the regressor. Comprehensive experiments are conducted on benchmarks FDDB and WIDER FACE to demonstrate the superiority of our proposed DSFD network compared with the state-of-the-art methods.

## References

- [1] Y. Abramson, B. Steux, and H. Ghorayeb. Yet even faster (yef) real-time object detection. *International Journal of Intelligent Systems Technologies and Applications*, 2(2-3):102–112, 2007. 2
- [2] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016. 2
- [3] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou. Selective refinement network for high performance face detection. In *Proceedings of Association for the Advancement of Artificial Intelligence (AAAI)*, 2019. 2, 7
- [4] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2016. 2
- [5] R. Girshick. Fast r-cnn. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2015. 2, 3
- [6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014. 2
- [7] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2
- [8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1
- [9] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical report, Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010. 7
- [10] K. Levi and Y. Weiss. Learning object detection from a small number of examples: the importance of good features. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004. 2
- [11] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2
- [12] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3, 4
- [13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 3
- [14] S. Liu, D. Huang, and Y. Wang. Receptive field block net for accurate and fast object detection. In *Proceedings of European Conference on Computer Vision*, 2018. 2, 4
- [15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *Proceedings of European conference on computer vision (ECCV)*, 2016. 2, 3, 4
- [16] W. Liu, A. Rabinovich, and A. Berg. Parsenet: Looking wider to see better. In *Proceedings of International Conference on Learning Representations Workshop*, 2016. 2
- [17] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2
- [18] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis. Ssh: Single stage headless face detector. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017. 2
- [19] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2015. 2, 4
- [20] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2014. 2
- [21] X. Tang, D. K. Du, Z. He, and J. Liu. Pyramidbox: A context-assisted single shot face detector. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2018. 2, 3, 5
- [22] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013. 2
- [23] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004. 1, 2
- [24] H. Wang, Z. Li, X. Ji, and Y. Wang. Face r-cnn. *arXiv preprint arXiv:1706.01061*, 2017. 2
- [25] J. Wang, Y. Yuan, and G. Yu. Face attention network: An effective face detector for the occluded faces. *arXiv preprint arXiv:1711.07246*, 2017. 3, 7
- [26] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen. Repulsion loss: Detecting pedestrians in a crowd. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3
- [27] Y. Wang, X. Ji, Z. Zhou, H. Wang, and Z. Li. Detecting faces using region-based fully convolutional networks. *arXiv preprint arXiv:1709.05256*, 2017. 2
- [28] S. Yang, P. Luo, C.-C. Loy, and X. Tang. Wider face: A face detection benchmark. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 7
- [29] C. Zhang, X. Xu, and D. Tu. Face detection using improved faster rcnn. *arXiv preprint arXiv:1802.02142*, 2018. 2
- [30] J. Zhang, X. Wu, J. Zhu, and S. C. Hoi. Feature agglomeration networks for single stage face detection. *arXiv preprint arXiv:1712.00721*, 2017. 2, 3
- [31] 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. 2

- [32] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. S<sup>3</sup>fd: Single shot scale-invariant face detector. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017. [2](#), [3](#), [5](#), [8](#)