DSFD: Dual Shot Face Detector

Jian Li†* Yabiao Wang‡ Changan Wang‡ Ying Tai‡
Jianjun Qian† Jian Yang† Chengjie Wang‡ Jilin Li‡ Feiyue Huang‡
†Nanjing University of Science and Technology ‡Youtu Lab, Tencent
[1]lijiannuist@gmail.com, {csjqian, csjyang}@njust.edu.cn
[2]{casewang, changanwang, yingtai, jasoncjwang, jerolinli, garyhuang}@tencent.com

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

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

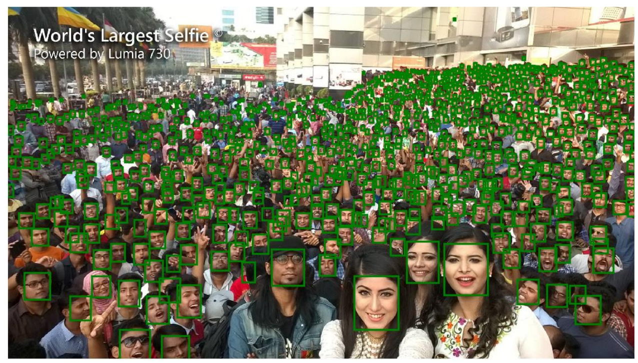


Figure 1: Our DSFD can find **861** faces out of the 1000 facial images present in the above image.

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 RCNN [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

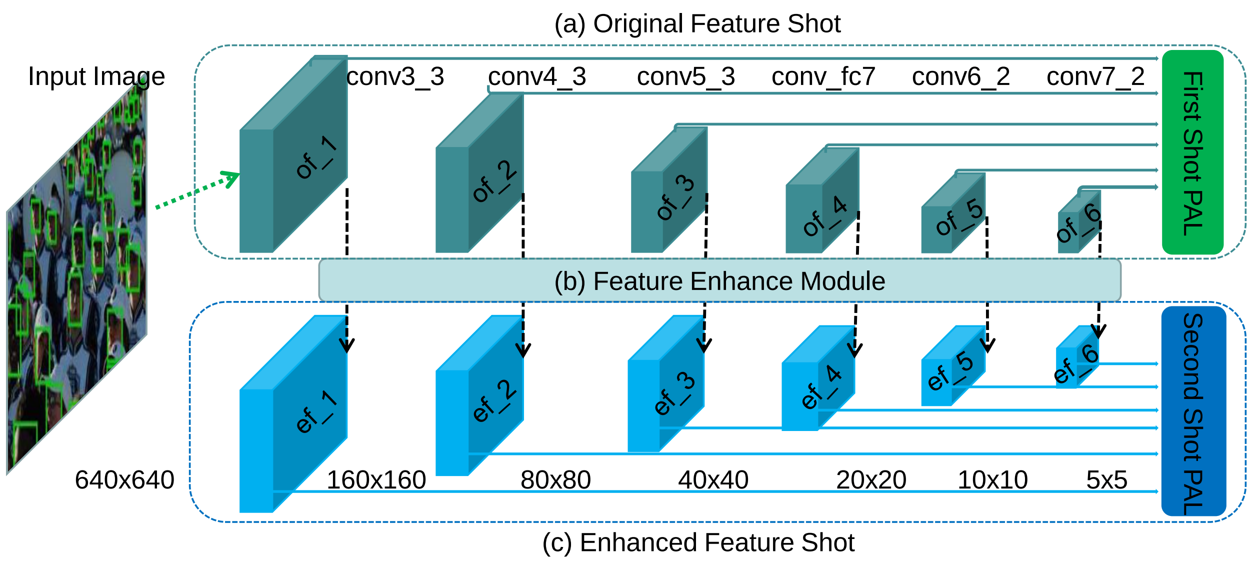


Figure 2: Our DSFD framework uses a Feature Enhance Module (b) on top of a feedforward VGG16 architecture to generate the enhanced features (c) from the original features (a), along with two loss layers named first shot PAL for the original features and second shot PAL for the enchanted features.

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 level.

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 prior works [21, 32] analyze a series of reasonable anchor scales and anchor compensation strategy to increase the number of positive anchors. However, such strategy ignores *random sampling* in data augmentation. Continuous face scale and a large number of discrete anchor scales still make the 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 features, since smaller anchor tiled to higher-level feature maps cell may have more semantic information for classification and high-resolution location information for detection. 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 that our smaller anchor tiling and improved anchor match with ground truth faces can eliminate the barrier of scale and occlusion to improve face detection performance.

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 segmentation. FPN [12], a top-down architecture, integrates 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] proposes smooth L_1 loss to prevent exploding gradients. Lin et al. [13] discovers 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] designs RepLoss for pedestrian detection, which improves performance in occlusion scenarios. FANet [30] creates 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] proposes an anchor compensation strategy to make tiny faces match enough anchors during training. Wang et al. [28] proposes a random crop strategy to generate large number of occluded faces for training. However, these methods ignore random sampling in data augmentation, while ours combines anchor assign strategy to provide better data initialization for anchor matching.

3. Dual Shot Face Detector

In this section, 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.

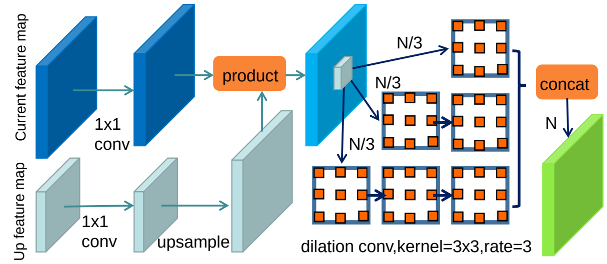


Figure 3: Feature Enhance Module illustrating the current feature map cell interactive with neighbor in current feature maps and up feature maps.

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 size as the original ones and are fed into SSD-style head to construct the second shot detection layers. Notice 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 PyramidBox, 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

The proposed feature enhance module aims to enhance the original features to make them more discriminate and robust. For a current anchor $a(i, j, l)$, FEM utilizes different dimension information including current layers anchors $a(i-1, j-1, l)$, $a(i-1, j, l)$, ..., $a(i, j+1, l)$, $a(i+1, j+1, l)$, and upper layers anchor $a(i, j, l+1)$. Specially, the feature map cell associated to anchor $a(i, j, l)$ can be mathematically defined as follow:

$$\begin{aligned} c_{i,j,l} &= f(\gamma(c_{i,j,l}), \delta(c_{i,j,l})) \\ \gamma(c_{i,j,l}) &= f(c_{i,j,l}, c_{i,j,l+1}) \\ \delta(c_{i,j,l})) &= f(c_{i-\varepsilon,j-\varepsilon,l}, c_{i-\varepsilon,j,l}, ..., c_{i,j+\varepsilon,l}, c_{i-\varepsilon,j+\varepsilon,l}) \end{aligned} \tag{1}$$

where $c_{i,j,l}$ is a cell located in (i, j) coordinate of the feature maps in the l-th layer, f denotes the combination of dilation convolution, element-wise product and up-sampling operation, and γ, δ denote current layer information and upper layer information, respectively. Fig. 3 illustrates the idea of FEM, which is inspired by FPN [12] and RFB [14]. Here, we first use 1×1 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 subnetworks 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:

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

where N is the number of matched dense boxes, 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. β is a weight to balance the effects of the two terms. 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:

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

and the two shots losses can be weighted summed into a whole Progressive Anchor Loss as follows:

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

Notice that anchor size in the first shot is half of ones in the second shot. 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.

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
of_1 (ef_1)	4	160×160	16 (8)	1.5 : 1	25600
of_2 (ef_2)	8	80×80	32 (16)	1.5 : 1	6400
of_3 (ef_3)	16	40×40	64 (32)	1.5 : 1	1600
of_4 (ef_4)	32	20×20	128 (64)	1.5 : 1	400
of_5 (ef_5)	64	10×10	256 (128)	1.5 : 1	100
of_6 (ef_6)	128	5×5	512 (256)	1.5 : 1	25

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

Component	Easy	Medium	Hard
S3FD+VGG16	92.6%	90.2%	79.1%
S3FD+VGG16+FEM	**93.0**%	**91.4**%	**84.6**%

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

Component	Easy	Medium	Hard
S3FD+RES50	93.7%	92.2%	81.8%
S3FD+RES50+FEM	95.0%	94.1%	88.0%
S3FD+RES50+FEM+PAL	**95.3**%	**94.4**%	**88.6**%

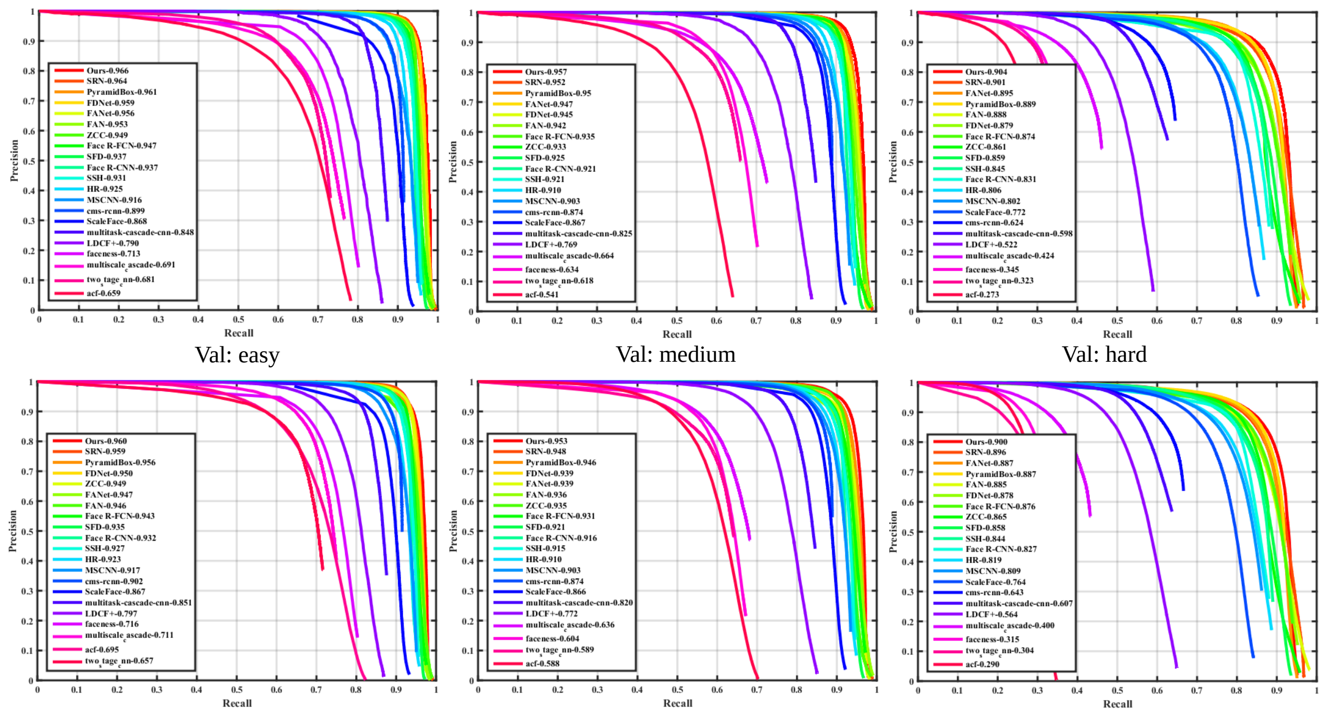
3.4. Improved Anchor Matching

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 to provide better initialization for the regressor.

Table 1 shows our anchor design details 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 used in SSD [15].

4. Experiments

In this section, we conduct extensive experiments and ablation studies to evaluate the effectiveness of the proposed framework. First, we demonstrate the impacts of several key components, including feature enhance module, progressive anchor loss, and improved anchor matching. Next, we compare our DSFD with the state-of-the-art face detectors on popular face detection benchmarks.



Val: easy | Val: medium | Val: hard

Test: easy | Test: medium | Test: hard

Figure 4: Precision-recall curves on WIDER FACE validation and testing subset.

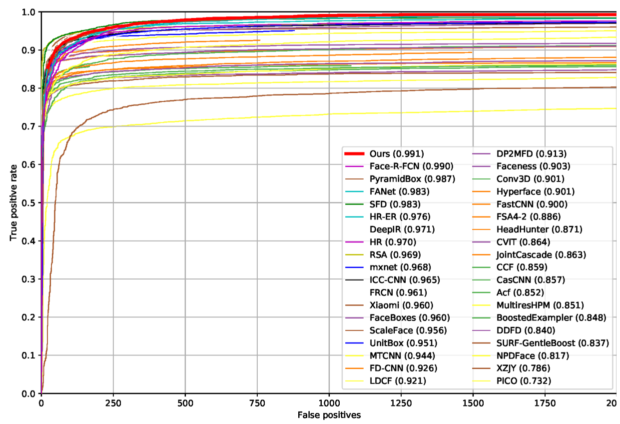
Table 4: Effectiveness of Improved Anchor Matching on the AP performance.

Component	Easy	Medium	Hard
S3FD+RES101	94.9%	93.3%	82.2%
S3FD+RES101+FEM+PAL	95.6%	94.7%	89.2%
S3FD+RES101+FEM+PAL+IAM	96.1%	95.3%	89.4%
S3FD+RES152+FEM+PAL+IAM+DeepHead+Expansion	**96.6**%	**95.7**%	90.4%
S3FD+RES152+FEM+PAL+IAM+DeepHead+Expansion+LargeBS	96.3%	95.5%	**91.0**%

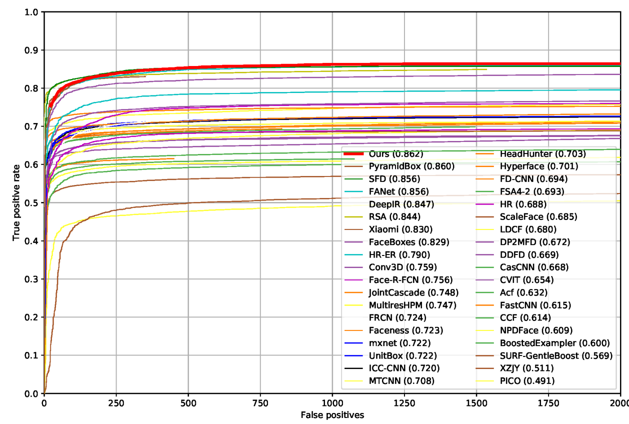
4.1. Analysis on DSFD

We conduct a set of ablation experiments on the WIDER FACE dataset to analyze our model in detail. For fair comparisons, we use the same parameter settings for all the experiments, 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. First, we use VGG16-based S3FD [32] without anchor compensation as the baseline. Table 2 shows that our feature enhance module can improve S3FD from 92.6%, 90.2%, 79.1% to 93.0%, 91.4%, 84.6%. Second, we use Res50-based S3FD scale-equitable framework without anchor compensation as the baseline. Table 3 shows our progressive anchor loss can improve Res50-based S3FD using FEM from 95.0%, 94.1%, 88.0% to 95.3%, 94.4%, 88.6%. To evaluate our improved anchor matching strategy, we use Res101-based S3FD without anchor compensation as the baseline. Table 4 shows that our APT can improve Res101-based S3FD using FEM and PAL from 95.6%, 94.7%, 89.2% to 96.1%, 95.3%, 89.4%.

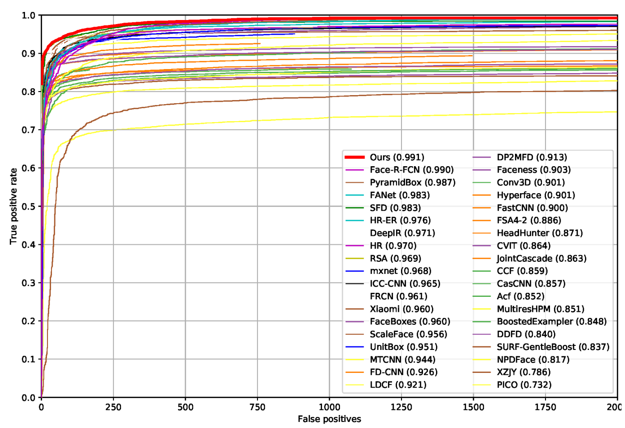
From the above three tables, some promising conclusions can be drawn: 1) Feature enhance is crucial. We use a more robust and discriminative feature enhance module to improve the feature presentation ability, especially for hard face. 2) Auxiliary loss based on progressive anchor is used to train all 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 0.5%, 0.6%, 0.2% on three settings, respectively. Moreover, we adopt several methods to further improve the detection performance, including expanding the detection coordinates (i.e., Expansion) and adding more convolutional layers for classification and regression branches simultaneously (i.e., DeepHead), which improves our DSFD to 96.6%, 95.7%, 90.4%. Additionally, when we enlarge the training batch



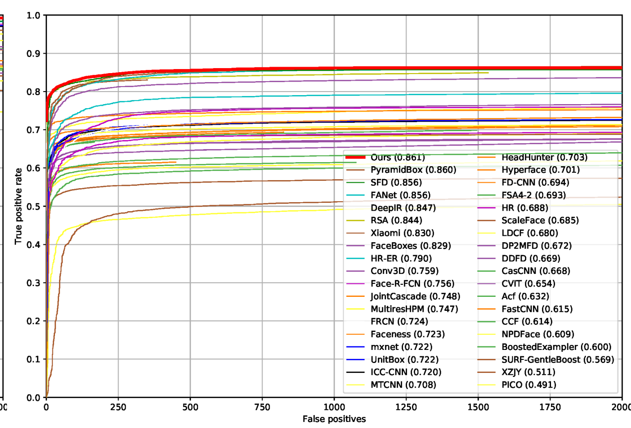
Discontinous ROC curves



Continous ROC curves



Discontinous ROC curves



Continous ROC curves

Figure 5: 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.

size (i.e., LargeBS), the result in hard setting can get 91.0% AP.

4.2. Comparisons with State-of-the-Art Methods

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

WIDER FACE Dataset It contains 393703 annotated faces with large variations in scale, pose and occlusion in total 32203 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 subset is further defined into three levels of difficulty: 'Easy', 'Medium', 'Hard' based on the detection rate of a baseline detector. As shown in Fig. 4, 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) for test set. Fig. 6 shows that our DSFD can handle faces with a wide range of face scales, especially for small faces. Green bounding boxes represent the detector confidence above 0.8. Fig. 7 shows that our DSFD is able to handle faces with various degrees of pose, occlusion, expression, appearance and illumination. These results demonstrate the effectiveness of our framework for finding faces with large variations.

FDDB Dataset It contains 5171 faces in 2845 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. 5, 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 1000. After adding additional annotations to those unlabeled faces [32], the false positives of our model can be further reduced and finally outperforms all other state-of-the-art methods.

Figure 6: Effectiveness of our DSFD on handling faces with a wide range of face scales. Green bounding boxes represent the detector confidence above 0.8.

Figure 7: Effectiveness of our DSFD on various large variations. Green bounding boxes represent the detector confidence above 0.8.

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] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *European Conference on Computer Vision (ECCV)*, 2016.

[3] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou. Selective refinement network for high performance face detection. *arXiv preprint arXiv:1809.02693*, 2018.

[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 (NIPS)*, 2016.

[5] R. Girshick. Fast r-cnn. In *Proceedings of IEEE International Conference on computer vision (ICCV)*, 2015.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[14] S. Liu, D. Huang, and Y. Wang. Receptive field block net for accurate and fast object detection. *arXiv preprint arXiv:1711.07767*, 2017.

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision (ECCV)*, 2016.

[16] W. Liu, A. Rabinovich, and A. Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[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.

[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.

[19] 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 (NIPS)*, 2015.

[20] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[21] X. Tang, D. K. Du, Z. He, and J. Liu. Pyramidbox: A context-assisted single shot face detector. In *European conference on computer vision (ECCV)*, 2018.

[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.

[23] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[24] H. Wang, Z. Li, X. Ji, and Y. Wang. Face r-cnn. *arXiv preprint arXiv:1706.01061*, 2017.

[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.

[26] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen. Repulsion loss: Detecting pedestrians in a crowd. *arXiv preprint arXiv:1711.07752*, 2017.

[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.

[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.

[29] C. Zhang, X. Xu, and D. Tu. Face detection using improved faster rcnn. *arXiv preprint arXiv:1802.02142*, 2018.

[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.

[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.

[32] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. Sˆ 3fd: Single shot scale-invariant face detector. In *Proceedings of IEEE International Conference on computer vision (ICCV)*, 2017.