Computer Systems Science & Engineering
DOI:10.32604/csse.2021.014086
Article

Tech Science Press

YOLOv3 Attention Face Detector with High Accuracy and Efficiency

Qiyuan Liu, Shuhua Lu* and Lingqiang Lan

College of Information and Cyber Security, People's Public Security University of China, Beijing, 102600, China
*Corresponding Author: Shuhua Lu. Email: lushuhual@ppsuc.edu.cn
Received: 30 August 2020; Accepted: 22 October 2020

Abstract: In recent years, face detection has attracted much attention and achieved great progress due to its extensively practical applications in the field of face based computer vision. However, the tradeoff between accuracy and efficiency of the face detectors still needs to be further studied. In this paper, using Darknet-53 as backbone, we propose an improved YOLOv3-attention model by introducing attention mechanism and data augmentation to obtain the robust face detector with high accuracy and efficiency. The attention mechanism is introduced to enhance much higher discrimination of the deep features, and the trick of data augmentation is used in the training procedure to achieve higher detection accuracy without significantly affecting the inference speed. The model has been trained and evaluated on the popular and challenging face detection benchmark, i.e., the WIDER FACE training and validation subsets, respectively, achieving AP of 0.942, 0.919 and 0.821 with the speed of 28FPS. This performance exceeds some existing SOTA algorithms, demonstrating acceptable accuracy and near real time detection for VGA resolution images, even in the complex scenarios. In addition, the proposed model shows good generation ability on another public dataset FDDB. The results indicate the proposed model is a promising face detector with high efficiency and accuracy in the wild.

Keywords: Face detection; YOLOv3; attention mechanism

1 Introduction

Face detection, as a significant branch of objective detection, has drawn much attention due to its extensively practical applications [1] in many human-machine interactive systems such as automation control, autofocus in commercial digital cameras, website login and verification etc. Meanwhile, face detection is the initial and basic step for a lot of face-based computer vision tasks including face recognition [2], face alignment [3], expression recognition [4] and so on.

During the past few decades, many efforts have been devoted to exploring the face detection techniques, resulting in remarkable progress [5,6,7,8,9,10]. It is the V-J face detector proposed by Viola et al. [5] that makes a breakthrough firstly in the face detection algorithms, which adopted Haar-Like feature and AdaBoost algorithm to train several cascaded binary classifiers, obtaining real time detection with acceptable accuracy. The V-J face detector and its improved counterparts [6,11,12,13] have good

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

performance on the front face detection. However, they suffer from low accuracy due to their intrinsic defects in feature extraction from the face images with the variation of scale, angle and occlusion. Another effective face detection method is called a deformable parts model (DPM) [14,15,16], proposed by P. Felzenszwalb in 2010s, which employs the improved histogram of oriented gradient (HOG) features and creates a combination between the sliding window algorithm and SVM classifier. DPM presents the multi-view face detection ability through describing the face as a series deformable parts connected with springs. Therefore, DPM based methods can detect faces in complex scenarios. On the contrary, their high computational cost makes detection speed very low, resulting in difficulty in the real-time application. Following AdaBoost and DPM methods, a series of the improved algorithms, such as MTM [17], HDPM [18] etc., have been developed for face detection. However, these approaches are based on the hand-crafted features, generally called the traditional face detectors, which shows non-robust performance in the wild scenarios owing to low flexibility and poor reliability.

Recently, with the rapid development of deep learning methods, convolutional neural network (CNN) based algorithms [19–25] have demonstrated amazing performance in face detection. That is, the deep learning methods present higher detection accuracy and stronger robustness than the traditional methods like AdaBoost and DPM mentioned above. Deep learning methods can automatically learn image deep features end to end through the network architecture, avoiding the subjective limitations of artificial designed features. Therefore, they exhibit the robust extraction ability applied into dealing with the unconstrained scenarios. Generally, deep learning methods for face detection are usually evolved from object detection. According to the feature processing approach, they can be roughly divided into two groups, named as one-stage detectors like YOLO [26–28], SSD [29], RetinaNet [30], and two-stage detectors like R-CNN [31], as well as its derivatives [32–34]. One-stage detectors have high detection speed but low accuracy due to their simple structure [28,35]. In contrast, two-stage detectors have good accuracy but low speed resulting from a large number of regions of interest (ROIs) through the proposal generator and refine procedure.

More recently, there are many state of the art (SOTA) face detectors including SSH [36], PyramidBox [22], Pyramidbox++ [25], DSFD [37], ASFD [38], RefineFace [39], EXTD [40], DFS [41], ISRN [24], FAN [42], RetinaFace [43] that have been employed to detect the unrestricted faces with severe occlusion, small size and variations in scale, pose and illumination etc. The SOTA methods aforementioned have been dominating the popular trend of the face detection conducted on GPU. Another face detectors [44,45], like Faceboxes [45], based on the CPU devices with low computational expense achieve promising average precision (AP) performance as well as real time speed in the WIDERFACE dataset [46], which may also be one of the representatives of face detection in the future. Additionally, it is worth mentioning that Chen et al. [47] proposed a YOLO face detector based on YOLOv3 Darknet-53, achieving real time detection speed up to 38 FPS with the accuracy of 69.3% on the WIDERFAC validation subset. Although the existing algorithms of face detection have made great progress, how to achieve the good tradeoff between accuracy and speed for face detection still needs to be further investigated in detail.

In the detection process, accuracy and speed are game relations, and what the study pursues is the good tradeoff. In this paper, we use Darknet-53 as the backbone and propose an improved YOLOv3 network with high accuracy and speed by introducing attention mechanism into the deep convolutional layers in order to enhance the deep feature distinguishability. We test our model on the WIDERFACE and FDDB datasets. The results demonstrate that the improvement is effective.

The main contributions of this paper are as follows: (1) the YOLOv3 model is highly improved for face detection by introducing attention mechanism and data augmentation. (2) The fine features and high-level semantics are realized to detect complex faces. (3) The proposed model achieves real time detection speed (28FPS) for VGA images with acceptable accuracy in the wild, which shows a better tradeoff

between efficiency and accuracy. (4) The proposed model presents good generation ability on another public dataset FDDB.

2 Methods

2.1 Framework

According to the previous reports [26,47,48], the YOLOv3 network has high detection speed, but relative low accuracy, especially for complex scenes and small size targets. In YOLOv3, Darknet-53 is not only more powerful than Darknet-19 but also more efficient than ResNet-101 and ResNet-152. Considering that, the architecture of the proposed model adopts Darknet-53 as the backbone and introduces attention mechanism and data augmentation, called the YOLOv3-attention model and shown in Fig. 1. Darknet-53 is employed for feature extraction, and the attention mechanism module is introduced to refine original feature maps to enhance discrimination of the deep features. In addition, the trick of data augmentation is used in the training procedure to achieve higher detection accuracy without significantly affecting the inference speed.

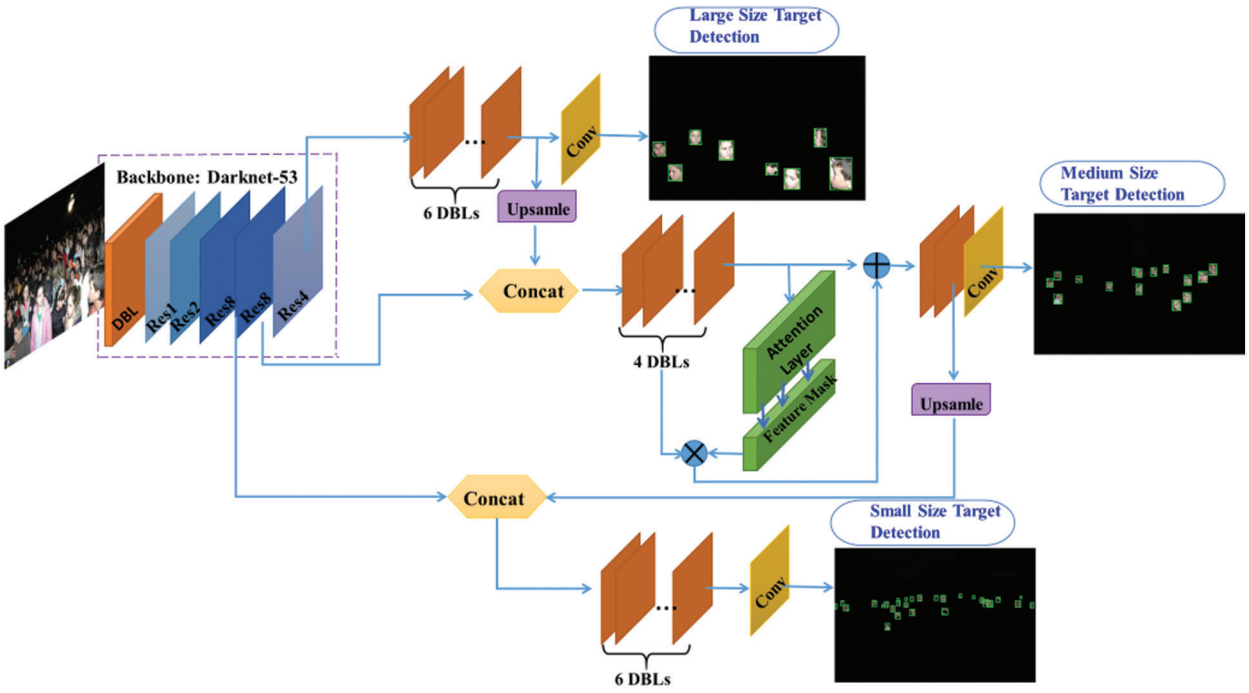

Figure 1: Architecture of the proposed YOLOv3-attention model. DBL, Darknetconv2d Batchnormalization Leaky

The detection process of YOLOv3 is carried out with 8, 16 and 32times down sampling, respectively, to realize the multi-scale feature detection, shown in Fig. 2. Taking the input picture with the 416 × 416 size and 3 channels as an example, a feature map of 26 × 26 is obtained by 16times down sampling (4 times 2× down sampling) at the 38th layer of the framework. Then, 32 times down sampling with a stride of 2 is conducted at the 63th layer, and the feature map scale turns to 1/32 of the input. After two step operations, the feature map has the largest receptive field, suitable for detecting larger targets. Subsequently, an up sampling operation with a stride of 2 and feature concatenation are executed to improve the nonlinearity and generalization ability of the network, suitable for detecting medium targets. Besides, that can reduce the parameters to

improve the real time performance. Finally, the obtained 8 times down sampling feature map has the smallest receptive field through repeating the operation of up sampling and feature concatenation at the 100th layer and 37th layer to further expand the feature dimension, achieving small target detection.

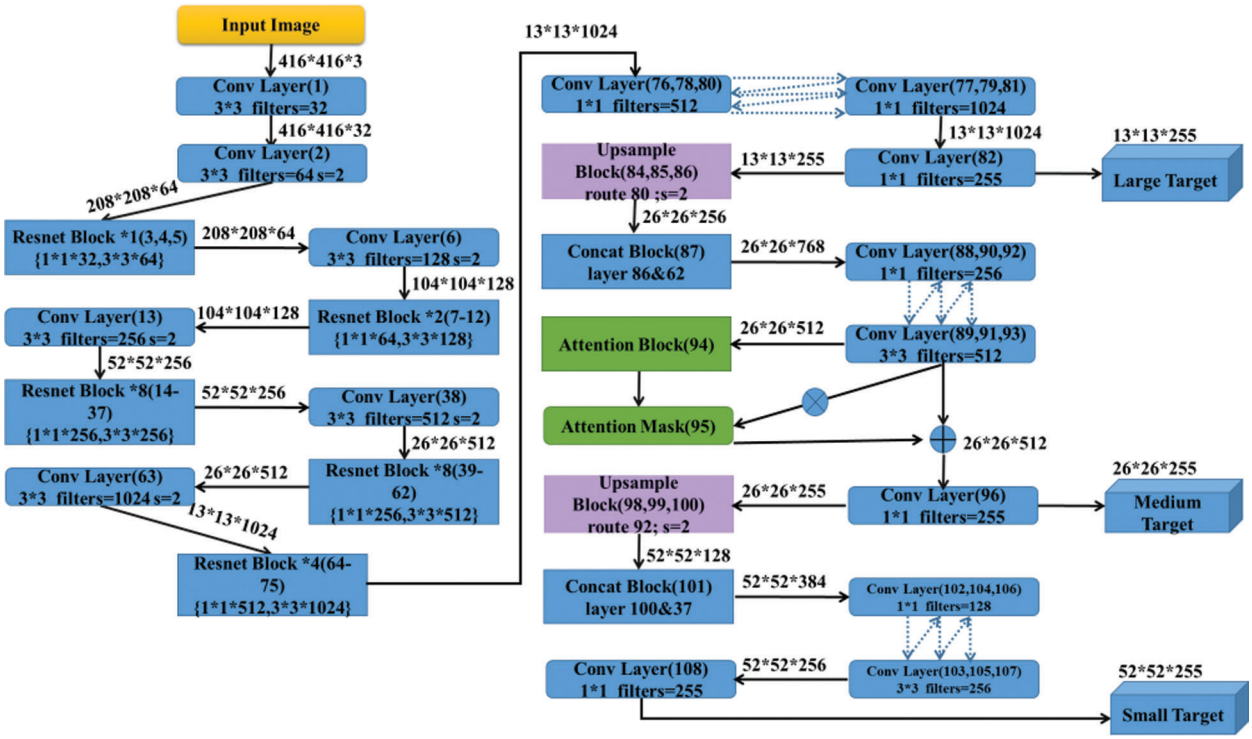

Figure 2: Convolution structure of the proposed YOLOv3-attention model

2.2 Attention Mechanism

The visual attention mechanism in deep learning is derived from the imitation of human vision, which can allocate information processing resources according to tasks. When humans process visual information, they often do not process and understand all the information. That is to say, they selectively pay attention to some significant or interesting information and automatically ignore other visible information. This helps to filter out some unimportant information and produce more distinguishing feature representation, resulting in improvement of the information processing efficiency [49].

In this paper, inspired by Song et al. [50], we introduce the soft attention module into the deep convolution layer to mask the original output. The results after the masking are multiplied by the original output and then superimposed back to the original output for the subsequent operation. In this way, we increase the distinguishability of effective information in the picture, and consequently improve the model detection ability.

With the deepening of the network layer, the feature noise is reduced, and the semantic information is enhanced. But there are still two problems: (1) deep features have low-resolution; (2) the ability of deep features to perceive details is poor. To solve the problem of low resolution, YOLOv3 adopts the way of up sampling and feature fusion with a stride of 2 when detecting medium and small size faces. That not only takes the advantages of deep features, but also enlarges the dimension of the feature map.

This article focuses on the second problem of YOLOv3 in deep feature learning. And so we seek to propose a deep feature enhance module (Fig. 1) in order to improve the ability to perceive the deep feature details. Accordingly, a soft attention module is designed between the 93th and 96th layers (Fig. 2) of the network, where the input as the previous convolution output is processed. And then the spatial features of the output are enlarged by the softmax layer. The attention layer generates the feature mask to re-weight the original input features with probability. The resulting feature mask is multiplied by the original output and then superimposed back to the original network as a new input to the next layer. The calculation process can be expressed as the following part.

The feature vector f_{ij} of the spatial location (i,j), and the mask of f_{ij} can be represented as follows,

$$mask_{i,j} = soft\max(score(f_{i,j}, \omega_{\partial})) \tag{1}$$

where, $mask_{i,j}$ is the mask probability of f_{ij} regularized by the attention layer, $score(\cdot)$ is the attention scoring mechanism based on different models that can be divided into addition, dot product and scaling dot product etc., and ω_{∂} is the matched weight of feature vector. Here, we use dot product scoring.

The mask dot products with the original feature output and gets the corresponding feature map $f_{i,j}^{att}$ of f_{ij}. The process can be expressed as follows,

$$f_{i,j}^{att} = mask_{i,j} \text{ e } f_{i,j} \tag{2}$$

where, the set of $f_{i,j}^{att}$ constitutes the attention map F^{att}, $F^{att} = \left\{f_{i,j}^{att}\right\}$, and the set of $mask_{i,j}$ constitutes the feature mask, $mask = \left\{mask_{i,j}\right\}$.

Considering that the introduction of attention mechanism will introduce some noise and omit some useful information in the process of reweighting, a shortcut connection is adopted to link the input of the attention module directly to its output by making an additive fusion. That is, the obtained attention map and the original input features are added together and then brought back to the original network as a new input for the subsequent layer. The final feature map F_{final}^{att} is depicted as,

$$F_{final}^{att} = \left\{(1 + mask_{i,j}) \text{ e } f_{i,j}\right\} \tag{3}$$

2.3 Data Augmentation

We propose a data augmentation strategy in order to prevent over-fitting, train a robust model, and as well as specially increase the detection effect of complex faces [51]. In the training set, the background part of the picture is used to randomly occlude a number of arbitrarily selected faces. Here, the number of occluded faces is controlled at half of the face index, which expands the original data and enhances the proportion of difficult samples.

3 Experiments

3.1 Datasets

3.1.1 Wider Face

WIDER FACE [46] is one of the largest and most challenging public face detection benchmarks. It contains 393703 annotated faces among 32203 images with a high degree of variability in scale, expression, pose and occlusion, etc. The dataset is randomly selected 40%, 10%, 50% images from 61 scene categories as the training, validation and testing subsets respectively. According to the detection results on the EdgeBox, the validation and test subsets are split into three difficulty levels, i.e., easy, medium and hard, for more diverse detection. WIDER FACE has the situations of label error, label noise and label omission, which will cause the model to stop training. We propose a data cleaning method to

eliminate the wrong data in it. In the training set, there are 12880 original pictures and 11612 correct pictures after cleaning.

3.1.2 FDDB

FDDB [52] is another challenging dataset widely used. The dataset is composed of 2,845 images with 5171 annotated faces collected from Yahoo News website, including a large amount of occlusion, low resolution, abnormal posture, etc. It provides two scoring methods: discrete score and continuous score. The scores of different algorithms are intuitively compared through receiver operating characteristic (ROC) curves.

3.2 Implementation Details

Our YOLOV3-attention model is trained on a NVIDIA GeForce GTX 2070 GPU and on two NVIDIA GeForce GTX 2080ti GPUs, respectively. The basic soft environment configuration is Cudnn7.4.0 and TensorFlow 1.13. On single 2070 GPU the batch size is set to 4 and the epochs are 200. The learning rate is 10^{-4} for the initial steps and with the epoch increasing, it exponentially decays to 10^{-7} finally. Anchor IoU threshold is 0.5. That means anchors of IoU > 0.5 will be assigned to positive classes. When trained on two 2080ti GPUs, the batch size is only set to 16 without other changes. The input size is set to 544 × 544 during the test time.

4 Results and Analysis

4.1 Accuracy and Analysis

To verify the effectiveness of the proposed model, we evaluate the results on the WIDER FACE validation subset. The precision-recall (PR) curves as well as AP scores of our proposed model (YOLOv3-attention) and the other existing methods are shown in Fig. 3. As can be seen from Fig. 3, our YOLOv3-attention model presents the AP of 0.942, 0.920, and 0.821, which is 4.8%, 4.5% and 5.3% higher than that of the YOLOv3 baseline (0.894, 0.874, and 0.768) on all the WIDER FACE validation easy, medium and hard subsets. This means that the attention mechanism introduced into the YOLOv3 network enhances the model's ability to perceive fine features and high-level semantics by concentration and fusion of features. Therefore, it results in effective improvement in the feature description. It is worth noting that our YOLOv3-attention model obtains the best improvement on the hard subset that contains large amount of complex faces such as large scale variances, occlusion and tiny faces etc. This demonstrates the effectiveness of the model for detecting complex faces, due mainly to the excellent learning ability and robustness.

In Fig. 3, it can also be seen that the YOLOv3-attention model is included in the SOTA class of the detection algorithms, especially on the easy and medium subsets. To conduct a quantitative comparison with the existing SOTA face detectors, we list the results in a summary Tab. 1. Among these SOTA face detectors, the proposed YOLOv3-attention model shows lower results but much higher efficiency (See 4.2 Section), when compared to some of two-stage detectors like PyramidBox [22] and PyramidBox++ [25] due to their proposal and refinement mechanism in processing. Meanwhile, our YOLOv3-attention model performs better than some latest face detectors like EXTD [40], SFD [53] in the easy or medium subset, but lower in the hard subset, resulting from the YOLOv3 simple structure. Besides, when compared to CMS-RCNN [21], MSCNN [54], RFAB-f-R-FCN [55] and YOLO-face-deeper darknet [47], our model has much higher AP across all the validation subsets, indicating the excellent performance. Objectively, it is worth mentioning that compared to some of SOTA face detectors, like ASFD [38], RefineFace [39] and RetinaFace [43], our model achieves lower results due to the lightweight network and fewer tricks. That may be improved by introducing more optimization strategies.

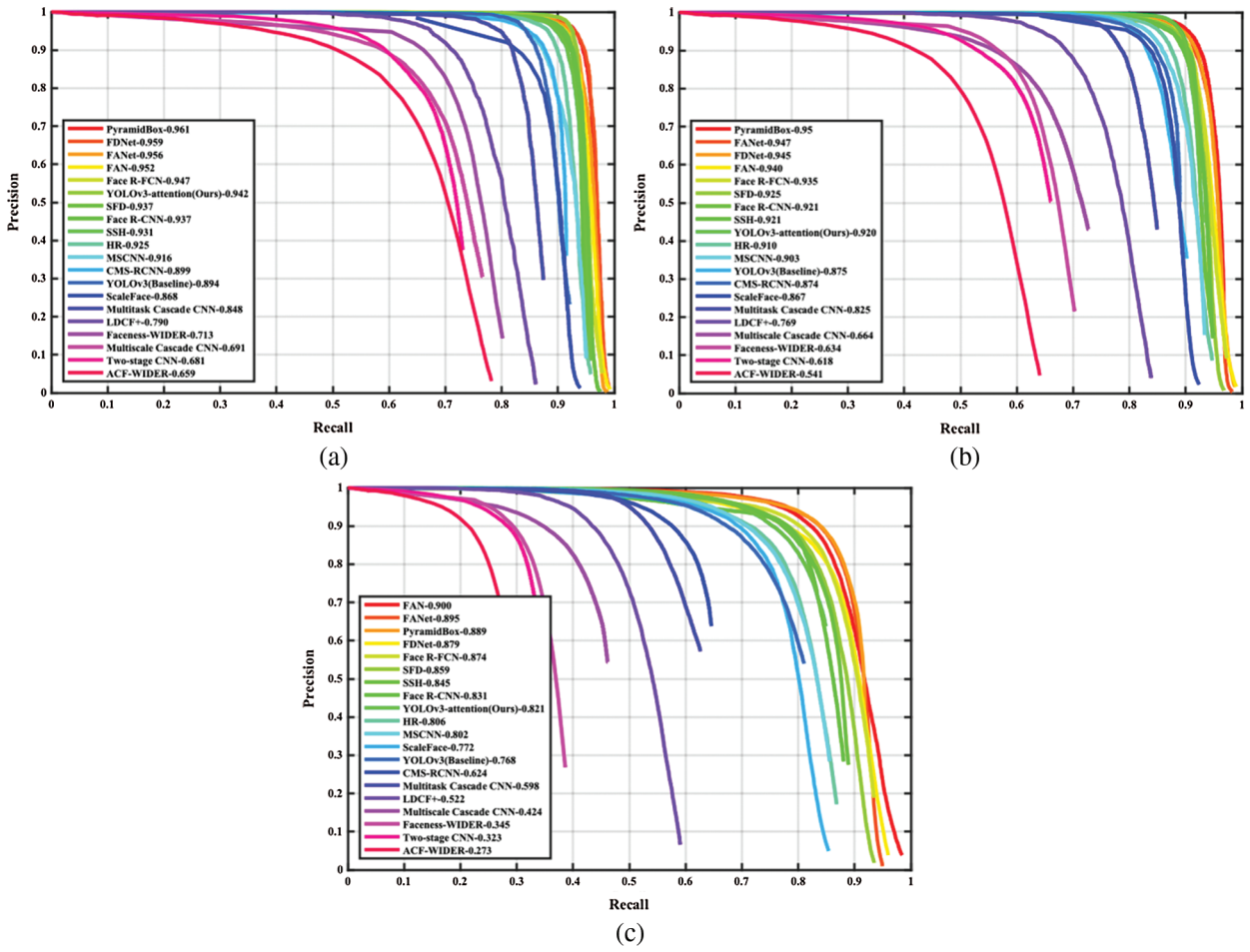

Figure 3: PR curves as well as AP scores of our proposed model on the WIDER FACE validation subset (a) Val-easy (b) Val-medium (c) Val-hard

Table 1: Comparison of our model with SOTA detectors on the WIDERFACE Val dataset

Model	Easy (AP)	Medium (AP)	Hard (AP)
ASFD-D6 [38]	0.972	0.965	0.925
RetinaFace(ResNet-152) [43]	0.969	0.961	0.918
Pyramidbox++ [25]	0.965	0.959	0.912
DSFD [37]	0.966	0.957	0.904
Pyramidbox [22]	0.961	0.950	0.889
FANet [56]	0.959	0.947	0.895
FAN [42]	0.952	0.940	0.900
Face R-FCN [57]	0.947	0.935	0.874
Face R-CNN [58]	0.937	0.921	0.831
EXTD-FPN-64-PReLU [40]	0.921	0.911	0.856

(Continued)

Table 1 (continued).

Model	Easy (AP)	Medium (AP)	Hard (AP)
EXTD-FPN-32-PReLU [40]	0.896	0.885	0.825
SFD [53]	0.937	0.925	0.859
SSH [36]	0.931	0.921	0.845
RFAB-f-R-FCN [55]	0.913	0.896	0.794
EagleEye [59]	0.863	0.792	0.462
MSCNN [54]	0.916	0.903	0.802
FaceBoxes [45]	0.885	0.862	0.773
CMS-RCNN [21]	0.899	0.874	0.624
YOLO-face-deeper darknet [47]	0.899	0.872	0.693
Faceness [60]	0.713	0.634	0.345
YOLOv3 (Baseline)	**0.894**	**0.874**	**0.768**
YOLOv3-attention (Ours)	**0.942**	**0.919**	**0.821**

To demonstrate the generalization ability, we test our proposed model trained on WIDER FACE on another classical face detection dataset FDDB. The discrete ROC curves are shown in Fig. 4. When the number of false positives is equal to 800, our YOLOv3-attention model gets the true positive rate of 96.5%. It outperforms RFAB-f-R-FCN [55], ICC-CNN [61], Faster-RCNN [32], FaceBoxes [45], etc., presenting good transfer inference ability.

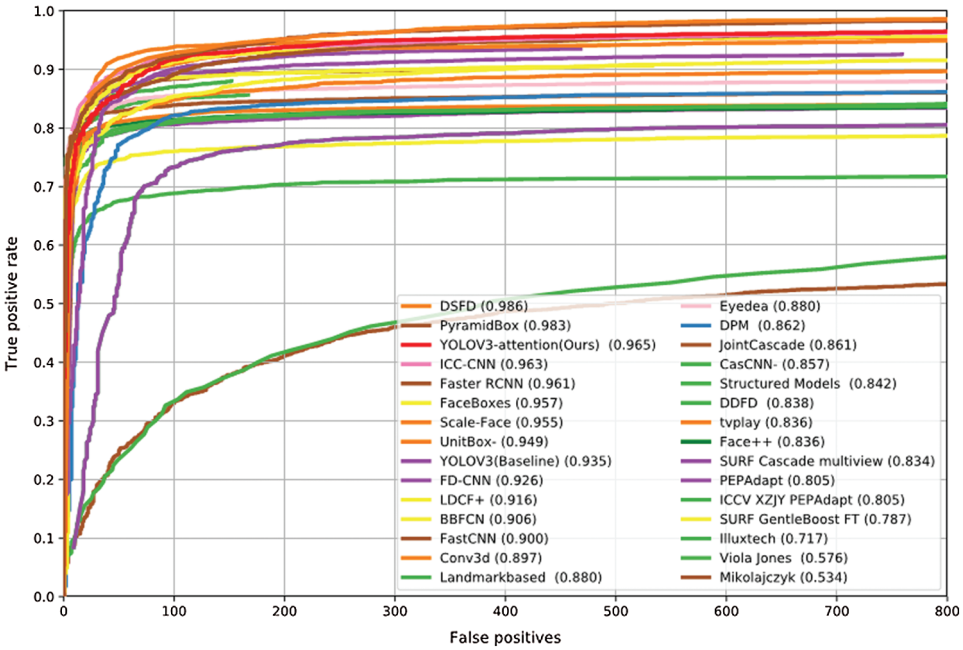

Figure 4: ROC curves on the FDDB dataset

To present the visualization detection effects of our proposed model, some experiment samples of YOLOv3 and its attention enhance model on the WIDERFACE dataset are shown in Fig. 5, respectively. It can be seen that the detection results (Fig. 5b) of our YOLOv3-attention model are 4 more faces than those (Fig. 5a) of YOLOv3. This demonstrates the proposed model is better than the baseline. In complex scenes, Fig. 5d detects 481 faces, far more than the 343 bounding boxes in Fig. 5c, especially for the small size and heavy occluded faces in the distance, indicating tremendous improvement in the recall rate via our enhance model. This can be attributed to the extraction of the discriminative deep features by an introduction of attention mechanism and data augmentation, presenting the face detector robustness in detecting tiny faces. In addition, it is interesting to note that Fig. 6 shows a qualitative result of the World Largest Selfie with dense faces under large variation in scale, occlusion, pose and blur. The proposed model detects around 900 faces out of the reported ~1100 faces even though including a few error cases. That is consistent with the detection of PyramidBox++ and RetinaFace, demonstrating the generalization and effectiveness of our proposed method in the wild.

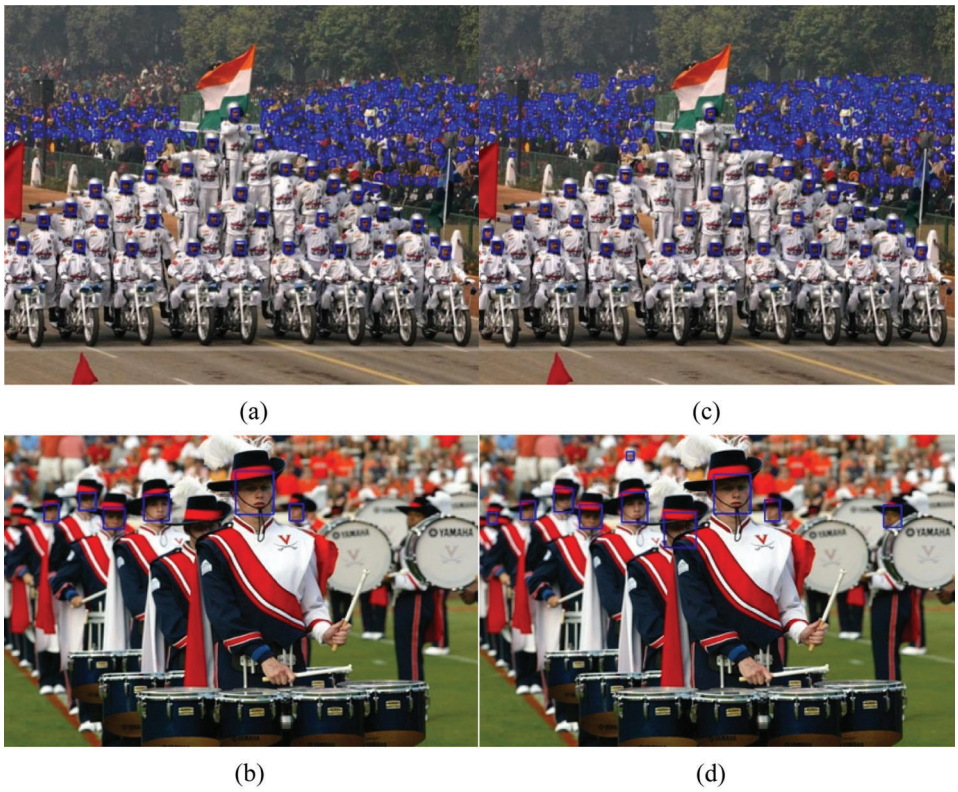

Figure 5: Some detection samples of YOLOv3 (a, b) and YOLOv3-attention model (c, d)

4.2 Speed and Analysis

As we know, it is of great importance for the efficiency of face detection in its practical applications, especially in some embedded devices. However, it is difficult to acquire the excellent tradeoff between efficiency and accuracy of face detectors. When the algorithm pursues the detection accuracy, it often needs a lot of calculation, i.e., complex computation procedures, resulting in low efficiency. To characterize the efficiency of our proposed model, we evaluate its inference speed results on the WIDERFACE validation subset, shown in Tab. 2. Tab. 2 also summarizes the speed results of the existing face detection models. As can be seen, the YOLOv3-attention and YOLOv3 models run 23FPS

and 23.7FPS on an ordinary 2070 GPU, respectively. This indicates that the deep feature enhancement module introduced into the network has almost no obvious effect on the detection speed. Furthermore, our YOLOv3-attention model runs 28 FPS on a 2080ti GPU with relative low cost under the 544 × 544 input image size, which shows a high efficiency face detector with near real-time speed (≥25FPS) detection for VGA resolution (640 × 480) images. As shown in Tab. 2, in terms of speed, our proposed model outperforms some of the SOTA face detectors such as PyramidBox [22], Face R-FCN [57] by a large margin. This can be attributed to the fact that two-stage methods have complex procedures and heavy weight structures. Certainly, compared to the YOLOv3 based face detector [47,51], our model demonstrates a little lower speed but much higher accuracy. For instance, the YOLO-face deeper darknet proposed by Chen et al. [47] has higher speed of 38FPS with the 416 × 416 input image size, but it has much lower accuracy of 69.3%. In addition, some of the SOTA face detectors with excellent accuracy and speed have been conducted on several GPUs with the high compute power [38,43]. Hence, our proposed model achieves a lot better tradeoff between efficiency and accuracy, as well as lower computation cost.

Figure 6: A detection example of the World Largest Selfie by YOLOv3-attention model

Table 2: Comparison of inference speed of face detectors on the WIDERFACE Val subset

Model	Speed/FPS	Hardware	AP
PyramidBox [22]	3	TITAN RTX GPU	(hard) 0.889
RetinaFace(ResNet-152) [43]	13	Tesla P40 GPU	(hard) 0.918
FAN-600 [42]	27.7	TITAN Xp GPU	(hard) 0.735
Face R-FCN [57]	3	Tesla K80 GPU	(hard) 0.874
YOLO-face deeper darknet [47]	38	GTX 1080ti GPU	(hard) 0.693
FaceBoxes [45]	10.2	CPU	(easy) 0.863
RFAB-f-R-RCN [55]	9.17	GTX 1080 GPU	(hard) 0.794
YOLOv3 (Baseline)	**23.7**	RTX 2070 GPU	(hard) **0.768**
YOLOv3-attention (Ours)	**23**	RTX 2070 GPU	(hard) **0.785**
YOLOv3- attention (Ours)	**28**	RTX 2080ti GPU	(hard) **0.821**

5 Conclusions

In this paper, using Darknet-53 as the backbone, we have proposed an improved face detector called the YOLOv3-attention model by introducing attention mechanism and data augmentation. The attention mechanism module is able to obtain more effective deep features containing fine-concentrated and high-level semantics by fusing the attention map and original input features. This results in enhancing much higher discrimination of the deep features. The data augmentation trick can construct a robust model and especially increase the detection effect of complex faces. Our proposed model has been evaluated on the WIDERFACE Val dataset, which achieves the AP of 0.942, 0.919and 0.821 with the speed of 28FPS. It demonstrates near real time detection with acceptable accuracy for VGA revolution images, indicating a better balance between efficiency and accuracy. In addition, the proposed model shows good generation ability on another public dataset FDDB. As a result, the proposed model is a promising face detector in the wild.

Funding Statement: This work is partially supported by National Key Research and Development Project (Grant No.A19808) and Fundamental Research Funds for the Central Universities (Grant No. 2019JKF225).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. Zafeiriou, C. Zhang and Z. Zhang, "A survey on face detection in the wild: Past, present and future," *Computer Vision and Image Understanding*, vol. 138, pp. 1–24, 2015.

[2] S. Li, F. Liu, J. Liang, Z. Cai and Z. Liang, "Optimization of face recognition system based on azure IoT edge," *CMC-Computers, Materials & Continua*, vol. 61, no. 3, pp. 1377–1389, 2019.

[3] J. Zhang, H. Hu and G. Shen, "Joint stacked hourglass network and salient region attention refinement for robust face alignment," *ACM Transactions on Multimedia Computing Communications and Applications*, vol. 16, no. 1, pp. 1–18, 2020.

[4] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1424–1445, 2000.

[5] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[6] J. Chen, S. Shan, S. Yang, X. Chen and W. Gao, "Modification of the adaboost-based detector for partially occluded faces," in *IEEE 2016 23rd Int. Conf. on Pattern Recognition*, pp. 516–519, 2016.

[7] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 99, pp. 1499–1503, 2016.

[8] Y. Xiao, D. Cao and L. Gao, "Face detection based on occlusion area detection and recovery," *Multimedia Tools and Applications*, vol. 79, no. 23–24, pp. 16531–16546, 2020.

[9] J. Yan, Z. Lei, L. Wen and S. Z. Li, "The fastest deformable part model for object detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2497–2504, 2014.

[10] Y. Li, B. Sun, T. Wu and Y. Wang, "Face detection with end-to-end integration of a convnet and a 3d model," in *The 14th European Conf. on Computer Vision*, pp. 420–436, 2016.

[11] S. Liao, A. K. Jain and S. Z. Li, "A fast and accurate unconstrained face detector," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 211–223, 2016.

[12] Y. Lin and T. Liu, "Robust face detection with multi-class boosting," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 680–687, 2005.

[13] P. Satyanarayana, N. J. Devi and S. K. S. Hasitha, "An enhanced Viola-Jones face detection method with skin mapping & segmentation," *Artificial Intelligence and Evolutionary Computations in Engineering Systems*, vol. 668, no. 1, pp. 485–493, 2018.

[14] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.

[15] P. F. Felzenszwalb, R. B. Girshick and D. McAllester, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[16] P. F. Felzenszwalb, R. B. Girshick and D. McAllester, "Cascade object detection with deformable part models," in *2013 IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2241–2248, 2013.

[17] D. Ramanan and X. Zhu, "Face detection, pose estimation, and landmark localization in the Wild," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2879–2886, 2012.

[18] G. Ghiasi and C. C. Fowlkes, "Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1899–1906, 2014.

[19] Y. Bai, Y. Zhang, M. Ding and B. Ghanem, "Finding tiny faces in the wild with generative adversarial network," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 21–30, 2018.

[20] J. Fu, S. R. Alvar, I. V. Bajic and R. G. Vaughan, "FDDB-360: Face detection in 360-degree fisheye images," in *IEEE Conf. on Multimedia Information Processing and Retrieval*, pp. 15–19, 2019.

[21] C. Zhu, Y. Zheng, K. Luu and M. Savvides, "CMS-RCNN: Contextual multi-scale region-based CNN for unconstrained face detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 57–79, 2016.

[22] X. Tang, D. K. Du, Z. He and J. Liu, "PyramidBox: A context-assisted single shot face detector," in *European Conf. on Computer Vision*, pp. 812–828, 2018.

[23] Y. Zhang, M. Ding and Y. Bai, "Detecting small faces in the wild based on generative adversarial network and contextual information," *Pattern Recognition*, vol. 94, pp. 74–86, 2019.

[24] S. Zhang, R. Zhu, X. Wang, H. Shi, T. Fu *et al.,* "Improved selective refinement network for face detection," 2019. [Online]. Available: https://arxiv.org/abs/1901.06651.

[25] Z. Li, X. Tang, J. Han, J. Liu and R. He, "Pyramidbox++: High performance detector for finding tiny face," 2019. [Online]. Available: https://arXiv:1904.00386.

[26] J. Redmon, S. Divvala and R. Girshick, "You only look once: Unified, real-Time object detection, " in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 779–788, 2016.

[27] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 7263–7271, 2017.

[28] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018. [Online]. Available: http://arXiv:1804.02767.

[29] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed *et al.,* "Single shot multibox detector," in *European Conf. on Computer Vision*, pp. 21–37, 2016.

[30] T. Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal loss for dense object detection," in *IEEE Int. Conf. on Computer Vision*, pp. 2980–2988, 2017.

[31] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 580–587, 2014.

[32] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2019. [Online]. Available: http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf.

[33] R. Girshick, "Fast R-CNN," in *IEEE Int. Conf. on Computer Vision*, pp. 1440–1448, 2015.

[34] K. He, G. Gkioxari, P. Dollar and R. Girshick, "Mask R-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386–397, 2020.

[35] S. Zhang, L. Wen, X. Bian, Z. Lei and S. Z. Li, "Single-shot refinement neural network for object detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 4203–4212, 2018.

[36] M. Najibi, P. Samangouei, R. Chellappa and L. S. Davis, "SSH: Single stage headless face detector," in *IEEE Int. Conf. on Computer Vision*, pp. 4885–4894, 2017.

[37] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian *et al.,* "Dual shot face detector," 2018. [Online]. Available: http://arXiv:1810.10220.

[38] B. Zhang, J. Li, Y. Wang, Y. Tai, C. Wang *et al.,* "ASFD: Automatic and scalable face detector," 2020. [Online]. Available: http://arXiv:2003.11228.

[39] S. Zhang, C. Chi, Z. Lei and S. Z. Li, "RefineFace: Refinement neural network for high performance face detection," 2019. [Online]. Available: http://arXiv:1909.04376.

[40] Y. Yoo, D. Han and S. Yun, "EXTD: Extremely tiny face detector via iterative filter reuse," 2019. [Online]. Available: http://arXiv:1906.06579.

[41] W. Tian, Z. Wang, H. Shen, W. Deng, B. Chen *et al.,* "Learning better features for face detection with feature fusion and segmentation supervision," 2018. [Online]. Available: http://arXiv:1811.08557.

[42] J. Wang, Y. Yuan and G. Yu, "Face attention network: An effective face detector for the occluded faces," 2017. [Online]. Available: http://arXiv:1711.07246.

[43] J. Deng, J. Guo, Y. X. Zhou, J. K. Yu, I. Kotsia *et al.,* "RetinaFace: Single-stage dense face localisation in the Wild," 2019. [Online]. Available: http://arXiv:1905.00641.

[44] C. Zhuang, S. Zhang, X. Zhu, Z. Lei, J. Wang *et al.,* "FLDet: A CPU real-time joint face and landmark detector," in *Int. Conf. on Biometric*, pp. 1–8, 2019.

[45] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang *et al.,* "Faceboxes: A CPU real-time face detector with high accuracy," in *2017 Int. Joint Conf. on Biometrics*, pp. 1–9, 2017.

[46] S. Yang, P. Luo, C. C. Loy and X. Tang, "Wider face: A face detection benchmark," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 5525–5533, 2016.

[47] W. Chen, H. Huang, S. Peng, C. Zhou and C. Zhang, "YOLO-face: A real-time face detector," 2020. [Online]. Available: https://doi.org/10.1007/s00371-020-01831-7.

[48] H. Zhang, W. Wei, X. Xiao, S. Yang and W. Shao, "Method to appraise dangerous class of building masonry component based on DC-YOLO model," *CMC-Computers, Materials & Continua*, vol. 63, no. 1, pp. 457–468, 2020.

[49] M. Luong, H. C. Pham and D. Manning, "Effective approaches to attention-based neural machine translation," 2015. [Online]. Available: http://arXiv:1508.04025.

[50] J. Song, Q. Yu and Y. Z. Song, "Deep spatial-semantic attention for fine-grained sketch-based image retrieval," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 5552–5561, 2017.

[51] B. Long, K. Yu and J. Qin, "Data augmentation for unbalanced face recognition training sets," *Neurocomputing*, vol. 235, no. 4, pp. 10–14, 2017.

[52] V. Jain and E. Learned-Miller, "FDDB: A benchmark for face detection in unconstrained settings," *UMass Amherst Technical Report, Technical Report UM-CS-2010-009*, vol. 2, no. 4, pp. 5–15, 2010.

[53] S. Zhang, X. Zhu, Z. Lei *et al.,* "S3FD: Single Shot Scale-invariant Face Detector," in *IEEE Int. Conf. on Computer Vision*, pp. 192–201, 2017.

[54] Z. Cai, F. Quan, R. S. Feris and N. A. Vasconcelos, "Unified multi-scale deep convolutional neural network for fast object detection," in *European Conf. on Computer Vision*, pp. 354–370, 2016.

[55] C. Tang, S. Chen, X. Zhou, S. Ruan and H. Wen, "Small-scale face detection based on improved R-FCN," *Applied Science*, vol. 10, no. 12, pp. 4177, 2020.

[56] J. Zhang, X. Wu and J. Zhu, "Feature agglomeration networks for single stage face detection," 2018. [Online]. Available: http://arXiv:1712.00721.

[57] Y. Wang, X. Ji, Z. Zhou, H. Wang and Z. Li, "Detecting faces using region-based fully convolutional networks," 2017. [Online]. Available: http://arXiv:1709.05256.

[58] H. Wang, Z. Li, X. Ji and Y. Wang, "Face R-CNN," 2017. [Online]. Available: http:// arXiv: 1706.01061.

[59] X. Zhao, X. Liang, C. Zhao, M. Tang and J. Wang, "Real-time multi-scale face detector on embedded devices," *Sensors*, vol. 19, no. 9, pp. 2158, 2019.

[60] S. Yang, P. Luo, C. C. Loy *et al.,* "From facial parts responses to face detection : A deep learning approach," in *IEEE International Conference on Computer Vision*, pp. 3676–3684, 2015.

[61] K. Zhang, Z. Zhang, H. Wang, Z. Li, Y. Qiao *et al.,* "Detecting faces using inside cascaded contextual CNN," in *Proc. of the IEEE Int. Conf. on Computer Vision*, pp. 3190–3198, 2017.