2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

Anchor Free Network for Multi-Scale Face Detection

Chengji Wang[1], Zhiming Luo[2], Sheng Lian[1] and Shaozi Li[1]
[1]Cognitive Science Department, Xiamen University, China
[2]Postdoc Center of Information and Communication Engineering, Xiamen University, China
Email: {chengji, zhiming.luo, lancerlian}@stu.xmu.edu.cn, szlig@xmu.edu.cn

***Abstract*—Anchor-based deep methods are the most widely used methods for face detection and have reached the state-of-the-art result. Compared with anchor-based methods that estimates the bounding-box rely on some pre-defined anchor boxes, anchor-free methods perform the localization by predicting the offsets of a pixel inside a face to its outside boundaries whose accuracies are much more precise. However, anchor-free methods suffer the drawback of low recall-rate mainly because 1) only using single scale features lead to miss detection of small faces, 2) the highly intra-class imbalance problem among different size faces. In this paper, to address these problems, we propose a unified anchor-free network for detecting multi-scale faces by leveraging the local and global contextual information of multi-layer features. We also utilize a scale aware sampling strategy to mitigate the intra-class imbalance issue which can adaptivity select the positive samples. Furthermore, a revised focal loss function is adopted to deal with the foreground/background imbalance issue. Experimental results on two benchmark datasets demonstrate the effective of our proposed method.**

I. Introduction

Human identity recognition is the key component of intelligent video analysis systems which have been widely used in access control, surveillance systems and other security applications. In a typical face recognition system, the first step is to localize the face in a given image which is called as face detection. As a special case of general object detection problem, CNNs based methods [5], [17] had achieved gratifying results on face detection problem. Current CNN based face detection methods can be categorized into two categories: Anchor-Based Methods and Anchor-Free Methods as shown in Fig. 1.

Anchor-Based Methods: As shown in Fig. 1(a), most CNN based detection methods like Fast-RCNN [4], Faster-RCNN [21], SSD [12] compute the bounding box of an object by regressing the offsets from a predefined anchor box. At the training phase, a smoothed-L1 loss was usually used to measure the disagreement between the estimated offsets and the offsets corresponded to the ground truth.

Anchor-Free Methods: Different from anchor-based methods, another series of methods directly output values corresponding with the position and the size of an object from a given point (see Fig 1b). Stand for by DenseBox [6] and UnitBox [29]. These models learn to directly predict offsets from bounding box vertexes to points in region of interest. The backbones of them are fully convolution neural networks [14].

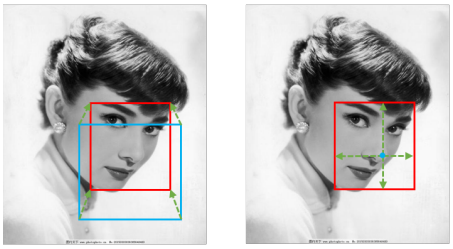

Fig. 1. A visual explanation shows the difference of anchor-based and anchor-free methods. The red bounding box is the ground truth, the blue bounding box is a predefined anchor, and the green lines are the offsets. (a) The anchor based methods predict the offsets based on predefined anchor. (b) The anchor-free methods directly estimate the offsets of a point to its outside boundaries.

With the state-of-the-art performance gained by anchor-based object detectors, as popularized in the Faster RCNN and SSD frameworks. Recently, most of the research attention about face detection are based on anchor-based methods, while anchor-free methods like DenseBox and UnitBox don't get very promising result on the large-scale WiderFace [27] challenge dataset. Two main drawbacks of anchor-free methods are as follows: First, the network architectures used by anchor-free methods don't use context information reasonably which then suffer the issue of detecting small objects. Second, excepting the inter-class imbalance problem, the intra-class imbalance problem also affects the performance of model. For example, the number of pixels of a big face with size 100×100 is sixteen times of a small face with size 25×25.

In this paper, we focus on above problems, leveraging anchor-free method to detect faces in unconstrained scene. As we known, small face detection is a challenging problem. Hu et al. [5] proved that incorporating context information can help to detect small faces, we address this by three mechanisms. The first one is multi-scale features fusion. Since the features from adjacent scales are complement with each other, a local fusion module is used to add local context information. Then a combination of multi-scale features is performed to increase receptive fields of small faces and maintain more detailed information of big faces. Secondly, we utilize scales related templates and intersection-over-union as metric to select positive examples within face region, which can balance the number of positive examples between big faces

978-1-5386-3788-3/18/$31.00 ©2018 IEEE

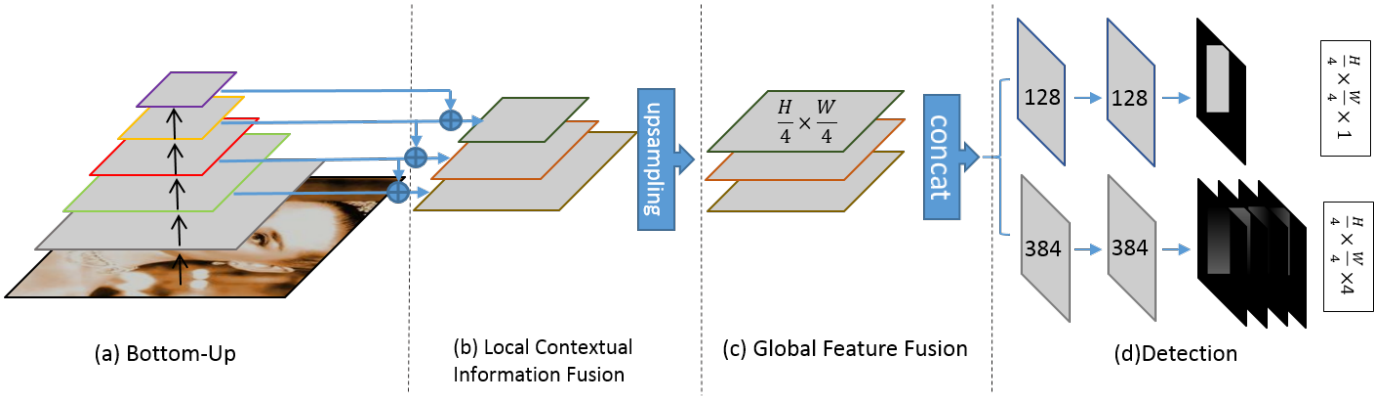

Fig. 2. The proposed detection system backbone on top of a feed-forward VGG architecture, combined with four major part: (a) Bottom-Up Pathway. (b) Local Contextual Information Fusion, merged local contextual information by addition. (c) Global Feature Fusion, adjust all features to the same resolution and merged in series. (d) Detection, based on fused feature we use two independent task related network to predict pixel level classification and bounding box regression.

and small faces. The last one, inspired by the focal loss [11], we propose a revised focal loss function to do hard examples mining which can further improve performance.

To sum up, the main contributions of this paper are as follows:

1) We proposed a multi-scale feature fusion framework for face detection, which efficiently fuse the local and global context information.
2) A scales aware sampling strategy is utilized to alleviate the intra-class imbalance problem.
3) A revised focal loss function is adopted to mining hard examples.

II. Related Work

In this section, we give a brief introduction of previous work on face detection. Based on the detection methodologies, we simple group previous work as following three categories:

Sliding-Window approaches: Classical detectors apply a classifier on a dense image grid which usually use hand-crafted features. Viola et al. [23] utilizes Haar-Like features and AdaBoost algorithm to train a cascade face/non-face classifier. Some works [25], [36] improve this paradigm by using more advanced features and classifiers, and [10], [30], [31] import CNN as feature extractors. Besides the cascade structure, [24], [37], [16] introduce deformable part models (DPM) [3] for face detection and achieve remarkable performance.

Anchor-Based methods: Inherited from the progress of generic anchor based object detection methods, Jiang et al. [8] applies Faster R-CNN [21] in face detection and get promising results. CMS-RCNN [35] also uses Faster R-CNN [21] in face detection by integrating body contextual information. Hu et al. [5] demonstrates that more anchor templates can improve recall rate as well as context information can help detecting small faces. Recently, [17], [33] follow the detection paradigm of SSD [12] and achieve the state-of-the-art performance. Qiao [19] try to predict object scales in images and Zhang [34] try to highlight object information on feature maps.

Anchor-Free methods: DenseBox[6] utilizes a fully convolutional neural network to regress a 4-D distance vector of each pixel inside a face (the offset of a pixel to the top, bottom, left and right boundaries of its candidate bounding box), and the L2 loss function is used for training. Since L2 loss function actually consider each offset in the 4-D distance vector independently, UnitBox [29] propose a new intersection-over-union loss function to jointly train this 4-D distance vector.

III. Proposed Method

In this section, we describe each component of our proposed method in detail. Firstly, we illustrate the overall architecture of our proposed network. Then we describe the definition of ground truth and the loss function used to train our network. Finally, we present a focal loss which is used to do hard example mining.

A. Network Architecture

As shown in Fig. 2, the proposed network is a single unified fully convolutional network that is consist of a feature extraction sub-network and a detection module. The feature extraction network contains a bottom-up pathway module, a local contextual information fusion module and a final global feature fusion module. The detection module has two sub-networks that are a classification sub-network and a bounding-box regression sub-network.

Bottom-Up Pathway: The bottom-up pathway is a feed forward backbone network, which computes a pyramid of feature maps at several scales with a scaling ratio of 2. In this paper, VGG-16 [22] that discarded all fully connected layers is used as the backbone. In order to capture multi-scale information in a single network, we select outputs from the pooling2, pooling3, pooling4 and pooling5 layers in the VGG-16 to construct the feature pyramid $\{P_2, P_3, P_4, P_5\}$. Notice that, this feature pyramid has strides of $\{4, 8, 16, 32\}$ with respect to the input image.

Local Contextual Information Fusion: Inspired by the work of Hu et al. [5] that suitable contextual information can improve detection result, we add several fusion blocks to

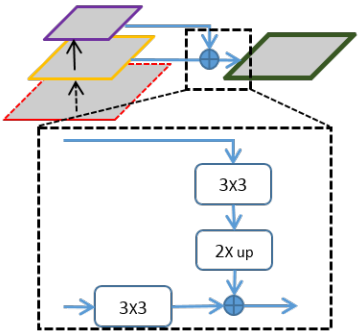

Fig. 3. A building block illustrate the local contextual feature fusion scheme.

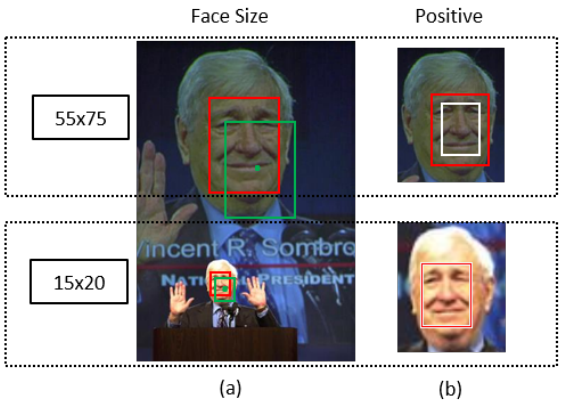

Fig. 4. An illustration of scale aware balance sampling, two faces are annotated with red box in the same image. Green boxes are templates in green point position. Pixels within white boxes are positive examples. The sizes are controlled by δ_{pos}.

embedded the local contextual information from higher layer into lower layer, as illustrated in Fig. 3.

Firstly, We restrict each scale's output to have the same feature channels by adding a 3×3 convolution layer respectively. Then an upsampling layer is followed to enlarge the spatial resolution of features from P_{n+1} as same as P_n. Finally, a element-wise summation is adopted to compute the local context fused feature map FL_n. The whole process is as follows:

$$FL_n = \text{Up}(\text{Conv}(P_{n+1})) + \text{Conv}(P_n) \quad (1)$$

where $n \in [2, 3, 4]$, "Up" is a upsampling layer, and "Conv" is a convolution layer with 128 filters.

Global Feature Fusion: Because high layer features encode more semantic information but with coarse spatial resolutions, which can not detect and localize small faces precisely. We use a global feature fusion module to fuse multi-layer features. As shown in Fig. 2(c), we first added two upsampling layers for the feature maps FL_4 and FL_3, then stacked together with FL_2 to get the final fused global features.

Detection: The detection module contains two parallel sub-networks, one for classification and another for bounding box regression. The classification sub-network adds three 3×3 convolutional layers with channels equal to 128, 128 and 1 on the top of global fused features, and the final output is a pixel probability map indicating whether a face or not. The regression sub-network has a similar structure while the feature channels are 384, 384 and 4, and the final outputs are 4 offset maps. The final output size of the network is 1/4 of original input image.

B. Loss Function

Following the paradigm of UnitBox [29], we also utilize a multi-task loss function to train our network. The full loss L can be represented as:

$$L = \lambda_{cls} \cdot L_{cls} + \lambda_{loc} \cdot L_{loc} + \lambda_{\Omega} \cdot \Omega(w) \quad (2)$$

where L_{cls} and L_{loc} are the loss of classification task and localization task respectively, $\Omega(w)$ denote regularization term over the network parameters w. Different loss terms are controlled by the hyper-parameter λ_{cls}, λ_{loc} and λ_{Ω}.

Classification task: The classification task can be deemed as a down-sampled segmentation task which classify each pixel into face and non-face. Instead of only consider all the pixels within a face as positives, we ignored some of pixels near the boundaries that not used for training. Then for an image, the groundtruth for each pixel is among one of the three categories {face, non-face, ignore} which are corresponded as $y_i^* \in \{1, 0, -1\}$.

The binary cross-entropy loss is used as the loss function L_{cls} for classification, denoted as:

$$L_{cls} = -\frac{1}{N} \sum_{i \in \{k | y_k^* \neq -1\}} (y_i^* \cdot \ln \widehat{y}_i + (1 - y_i^*) \cdot \ln(1 - \widehat{y}_i)) \quad (3)$$

where $\widehat{y}_i$ is predicted probablity for pixel i, and N is the size of set $\{k | y_k^* \neq -1\}$.

Localization task: The coordinate of a target bounding box can be represented by its left-top point $p_t = \{x_t, y_t\}$ and right-bottom point $p_b = \{x_b, y_b\}$. For each inside pixel (x_i, y_i), we can compute the position of its bounding box by a 4-dimensional offsets vector $\hat{t}_i = \{\hat{\mathrm{d}x}^t = |x_i - x_t|, \hat{\mathrm{d}y}^t = |y_i - y_t|, \hat{\mathrm{d}x}^b = |x_i - x_b|, \hat{\mathrm{d}y}^b = |y_i - y_b|\}$ which are the distances to the four boundaries. The goal for our localization task is to estimated this offsets as accurate as possible. In order to achieve this goal, we choose the IoU Loss [29] as the loss function of the localization sub-network. The localization loss is defined as follow:

$$L_{loc} = -\frac{1}{N} \sum_{i \in \{k | y_k^* = 1\}} \ln(iou_i) \quad (4)$$

and

$$iou_i = \text{IoU}(\hat{R}_i, R_i^*) = \frac{|\hat{R}_i \bigcap R_i^*|}{|\hat{R}_i \bigcup R_i^*|} \quad (5)$$

where $\hat{R}_i$ is the predicted box of pixel i and R_i^* is its corresponding ground truth, N is the size of set $\{k | y_k^* = 1\}$. This loss is only computed on those positive examples.

C. Scale aware sampling

As showed in Figure 4, the size of big face is over ten times of the small face. In a result, the spatial size of feature maps related to small face is less than $1/10$ of big faces, that leads a bias in the model during training. In order to alleviate this problem, we proposed a scale aware positive example sampling strategy by implicitly incorporate positive samples selection methodology in anchor-based methods.

For a pixel p_j inside a face region, the ground truth box is G_j (red box in Fig. 4) and a template T_j with the same size of G_j centered at p_j. We utilize the IoU of G_j and T_j as a metric, and set pixel p_j as positive example if the IoU is above a predefined threshold.

$$l_j = \begin{cases} 1, & \text{if IoU}(G_j, T_j) > \delta_{pos} \\ -1, & \text{otherwise} \end{cases} \tag{6}$$

In this paper, we set $\delta_{pos} = 0.5$ for big faces and $\delta_{pos} = 0.2$ for small faces whose size are small than 25×25.

D. Focal Loss

The original focal loss [11] is proposed to do hard example mining during training by adaptive adjusting the weights of different samples based on their difficulties, but it is only used for classification task. In this paper, we extend focal loss to do hard example mining on both classification and localization task. For the classification task, the weights of pixels with higher probabilities will be decreased which mean they are easy to be classified. For the localization task, we will lower the weights of those pixels with higher IoU with the groundtruth. The final focal loss for both tasks are as follows:

Classification task: we use the same formula to compute the weights as in [11].

$$prob_i = \begin{cases} \hat{y_i}, & \text{if } y_i^* = 1 \\ 1 - \hat{y_i}, & \text{otherwise} \end{cases} \tag{7}$$

$$L_{cls} = -\frac{1}{N} \sum_{i \in \{k | y_k^* = 1\}} (1 - prob_i)^2 \ln(prob_i) \tag{8}$$

Localization task: we regard the IoU of a estimate bounding box with its groundtruth as a probability, and set the weight as $(1 - iou_i)$.

$$L_{loc} = -\frac{1}{N} \sum_{i \in \{k | y_k^* = 1\}} (1 - iou_i) \ln(iou_i) \tag{9}$$

IV. Experiments

The experiment section is organized as following: we first introduce the basic datasets and experiment setup, then a model analysis is conducted to verify the effectiveness of each component in our method, finally we compare our method with state-of-the-art on two benchmarks.

A. Datasets and Experiment Setup

The FDDB [7] and WIDER Face [27] benchmark datasets have been used to evaluation our method. The FDDB dataset contains 2,845 images with 5,171 annotated faces. The WIDER Face dataset has 32,203 images with 393,703 labeled faces, and 50% of the images are used for testing, 40% for training and the remaining for validation. In this paper, all the models are trained on WIDER Face training set without any data augmentation, we report the results on the FDDB and WIDER Face validation set.

As shown in Fig. 2(d), the sigmoid function is used to normalize the outputs of the classification sub-network to a probability in the range [0,1] and the ReLU is adopted as the activation function for the regression sub-network. During the training phase, we only use one image per-batch with size 1024×1024 due to the limitation of GPU memory. The random crop and padding are used to generate the training images. The Adam optimizer [9] is utilized to train our model for 40 epochs with an initial learning rate $1e^{-6}$, a weight decay 0.0005. λ_{loc} is set to 1 and λ_{cls} is set to 0.01. For the testing, we only test in single scale and use a non maximal suppression (NMS) with a threshold of 0.45 to generate the final results.

B. Model Analysis

We analyze the effectiveness of each component in our method by extensive experiments on the Wider Face validation set and the FDDB dataset. The Wider Face validation set has easy, medium and hard subsets, which can be roughly considered as large, medium and small faces.

Baseline: In order to evaluate the feature fusion strategy of our model, we trained a face detector with the same network architecture of UnitBox [29] as our baseline. This architecture doesn't have any feature fusion which only use the features of pooling4 to compute the face probability and the features of pooling5 to do the bounding box regression. For this baseline, we use the same training settings but without scale aware balance sampling and focal loss strategy.

Ablative Setting: To examine how each component affects the final performance, we do a ablation experiment of our model under three different settings. (1) *Feature Fusion*: the model is trained only use the architecture in Fig. 2 without another two components. (2) *Scale aware sampling*: the model is trained with the balance sampled positive examples. (3) *Focal Loss*: this is our complete model, consisting the *scale aware sampling* and *focal loss*.

The results on the Wider Face validation set are reported in TABLE I. Compared with baseline model, the feature fusion strategy can give significant boost of mean-average-precision (MAP). Especially on the *Hard* subset of Wider Face, our feature fusion network improve MAP over 50% which means enhancing the low layer features by the context information from the high layer features is essential for the detection of small face. By adding the scale aware sampling, we improvement the MAP around 0.01 on the *Easy* and *Medium* subsets while with slightly decrease on the *Hard* subset. Our complete model contains all the components can further increase the

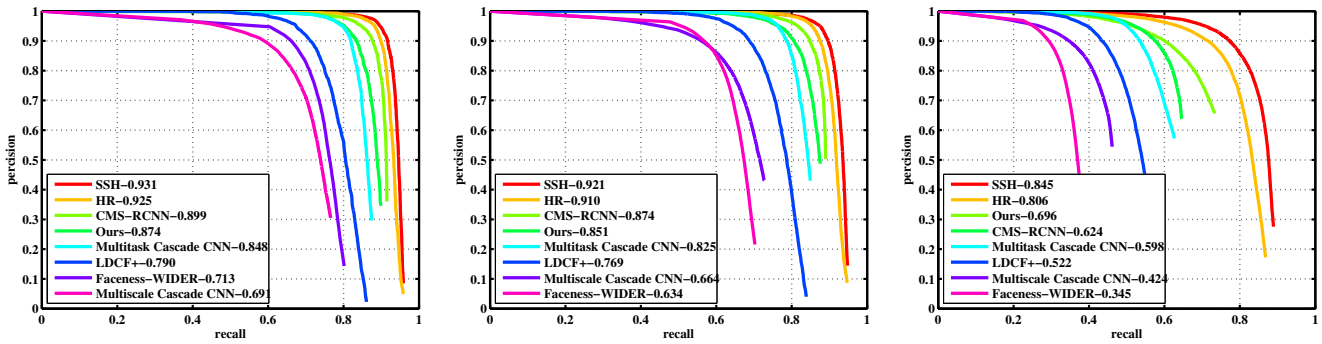

Fig. 5. Precision-Recall curves on Wider Face validation set.

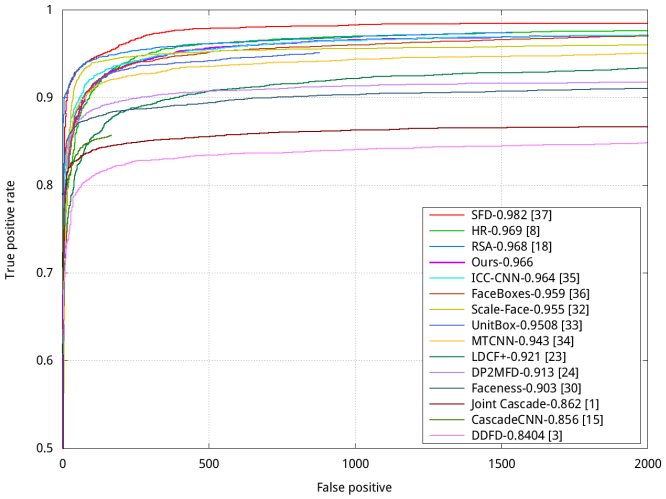

Fig. 6. Compared with state-of-the-art by using discontinuous ROC curves on FDDB dataset.

MAP for all three subsets. For the FDDB dataset, we can get a similar performance as shown in TABLE II.

TABLE I
THE MEAN-AVERAGE-PRECISION (MAP) ON WIDER FACE VALIDATION SET.

Methods	Easy	Medium	Hard
Baseline	0.757	0.737	0.404
Feature Fusion	0.833	0.797	0.626
Scale Aware Sampling	0.846	0.805	0.623
Focal Loss	0.874	0.851	0.696

TABLE II
TRUE POSITIVE RATE ON FDDB FOR 1000 FALSE POSITIVES

Baseline	Feature Fusion	Scale Aware Sampling	Focal Loss
0.906	0.946	0.956	0.966

C. Evaluation on benchmark

FDDB dataset: We compare our face detector by using the FDDB evaluation protocol with the state-of-the-art methods [33], [2], [18], [28], [31], [15], [30], [32], [5], [29], [20], [26], [13], [1], [10]. We plot the discontinuous ROC curves in Fig. 6 with the results of other methods downloaded from FDDB official site[1]. Our method outperform most of the state-of-the-art methods which indicate that our method can be used to detect faces in unconstrained scenes.

Wider Face: Faces in the Wider Face dataset are divided into three levels (Easy, Medium and Hard subset) according to the difficulties of the detection. We compare our method with several recent state-of-the-art [5], [17], [30], [18], [27], [2], [26] on the validation set. The Precision-Recall curves of the three levels is plot in Fig. 5, and some examples of detection result of our method is shown in Fig. 7.

As can be seen from Fig. 5 and Fig. 6, some of the recent anchor-based methods have better performance than our method. There are still lots of area we need to explored to further improve our method.

V. CONCLUSION

In this paper, a novel face detector is proposed to deal with the performance degeneration of detecting small faces in common anchor-free methods. After analyzing the issue of this phenomenon, we design a multi-scale feature fusion framework which combine local contextual information from a wide range of layers that it is very useful for detecting small faces. Besides, a scale aware example sampling strategy is used to improve the recall rate of detection which successfully import template in anchor-free detection paradigm. For the future work, we intend to further improve the classification strategy of background patches, and explore the relation of IoU Loss and templates.

VI. ACKNOWLEDGEMENTS

This work is supported by the National Nature Science Foundation of China (No. 61572409, No. U1705286 & No. 61571188), the Natural Science Foundation of Jiangxi Province (20171BAB212013), Fujian Province 2011 Collaborative Innovation Center of TCM Health Management and Collaborative Innovation Center of Chinese Oolong Tea IndustryCollaborative Innovation Center (2011) of Fujian Province.

[1]http://vis-www.cs.umass.edu/fddb/results.html

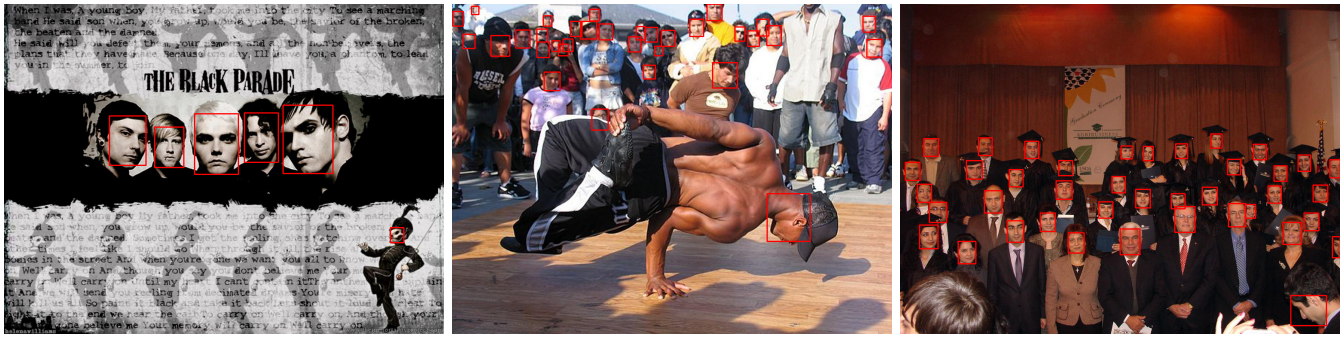

Fig. 7. Examples of detection result on Wider Face.

REFERENCES

[1] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, "Joint cascade face detection and alignment," in *European Conference on Computer Vision*. Springer, 2014, pp. 109–122.

[2] S. S. Farfade, M. J. Saberian, and L.-J. Li, "Multi-view face detection using deep convolutional neural networks," in *ACM Int. Conf. Multimedia Retrieval*, 2015, pp. 643–650.

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, 2010.

[4] R. Girshick, "Fast R-CNN," in *Proc. ICCV*, 2015, pp. 1440–1448.

[5] P. Hu and D. Ramanan, "Finding tiny faces," in *Proc. CVPR*, 2017, pp. 1522–1530.

[6] L. Huang, Y. Yang, Y. Deng, and Y. Yu, "Densebox: Unifying landmark localization with end to end object detection," *arXiv preprint arXiv:1509.04874*, 2015.

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

[8] H. Jiang and E. Learnedmiller, "Face detection with the Faster R-CNN," in *IEEE Int. Conf. Automatic Face Gesture Recognition*, 2017, pp. 650–657.

[9] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[10] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proc. CVPR*, 2015, pp. 5325–5334.

[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *arXiv preprint arXiv:1708.02002*, 2017.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. ECCV*, 2016, pp. 21–37.

[13] Y. Liu, H. Li, J. Yan, F. Wei, X. Wang, and X. Tang, "Recurrent scale approximation for object detection in cnn," in *IEEE International Conference on Computer Vision*, 2017.

[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, 2015, pp. 3431–3440.

[15] N. Markuš, M. Frljak, I. S. Pandžić, J. Ahlberg, and R. Forchheimer, "Object detection with pixel intensity comparisons organized in decision trees," *arXiv preprint arXiv:1305.4537*, 2013.

[16] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *Proc. ECCV*, 2014, pp. 720–735.

[17] M. Najibi, P. Samangouei, R. Chellappa, and L. Davis, "Ssh: Single stage headless face detector," in *Proc. CVPR*, 2017, pp. 4875–4884.

[18] E. Ohn-Bar and M. M. Trivedi, "To boost or not to boost? on the limits of boosted trees for object detection," in *Proc. ICPR*, 2016, pp. 3350–3355.

[19] S. Qiao, W. Shen, W. Qiu, C. Liu, and A. Yuille, "Scalenet: Guiding object proposal generation in supermarkets and beyond," in *ICCV*, 2017.

[20] R. Ranjan, V. M. Patel, and R. Chellappa, "A deep pyramid deformable part model for face detection," in *IEEE Int. Conf. Biometrics Theory, Applications and Systems (BTAS)*, 2015, pp. 1–8.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NIPS*, 2015, pp. 91–99.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[23] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[24] J. Yan, Z. Lei, L. Wen, and S. Z. Li, "The fastest deformable part model for object detection," in *Proc. CVPR*, 2014, pp. 2497–2504.

[25] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Aggregate channel features for multi-view face detection," in *IEEE Int. Joint Conf. Biometrics (IJCB)*, 2014, pp. 1–8.

[26] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *Proc. ICCV*, 2015, pp. 3676–3684.

[27] ——, "Wider face: A face detection benchmark," in *Proc. CVPR*, 2016, pp. 5525–5533.

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

[29] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "Unitbox: An advanced object detection network," in *Proc. ACM Multimedia*, 2016, pp. 516–520.

[30] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, 2016.

[31] K. Zhang, Z. Zhang, H. Wang, Z. Li, Y. Qiao, and W. Liu, "Detecting faces using inside cascaded contextual cnn," in *Proc. CVPR*, 2017, pp. 3171–3179.

[32] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "Faceboxes: a cpu real-time face detector with high accuracy," *arXiv preprint arXiv:1708.05234*, 2017.

[33] ——, "S3fd: Single shot scale-invariant face detector," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 192–201.

[34] Z. Zhang, S. Qiao, C. Xie, W. Shen, B. Wang, and A. L. Yuille, "Single-shot object detection with enriched semantics," *arXiv preprint arXiv:1712.00433*, 2017.

[35] C. Zhu, Y. Zheng, K. Luu, and M. Savvides, "CMS-RCNN: contextual multi-scale region-based cnn for unconstrained face detection," in *Deep Learning for Biometrics*. Springer, 2017, pp. 57–79.

[36] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Proc. CVPR*, vol. 2, 2006, pp. 1491–1498.

[37] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. CVPR*, 2012, pp. 2879–2886.