Dense Face Detection via High-level Context Mining

Qixiang Geng[1], Dong Liang[1], Huiyu Zhou[2], Liyan Zhang[1], Han Sun[1] and Ningzhong Liu[1]
[1] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
MIIT Key Laboratory of Pattern Analysis and Machine Intelligence
Collaborative Innovation Center of Novel Software Technology and Industrialization
[2] School of Informatics, University of Leicester
{gengqx, liangdong, sunhan, zhangliyan}@nuaa.edu.cn hz143@leicester.ac.uk lnz_nuaa@163.com

2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) | 978-1-6654-3176-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/FG52635.2021.9667044

***Abstract*— The appearance degradation caused by low resolution is the core problem of small face detection. Therefore, a natural approach is to assemble information from the context. This paper focuses on how to use high-level contextual information to improve the abilities of anchor-based detectors to detect dense and degenerate faces. We tap the spatial contextual information on the overall view based on the density map, and propose the prior of face co-occurrence for inferred bounding-boxes coordination. We also propose score-size-specific non-maximum suppression to replace the traditional non-maximum suppression at the end of anchor-based detectors. According to the inferred face boxes' quantity, score and size, the proposed synthetical solution reduces false positives and increases true positives. Our method does not require additional training, which is model-independent and can be embedded into existing face detectors. We also propose a dataset - Crowd Face for face detection, which is full of challenges. We expect to supply enough samples to highlight the difficulties of detecting dense and degenerate faces. We embed our proposed methods into state-of-the-art face detectors on massively benchmarked face datasets. Compared with the prior art on the WIDER FACE hard set, our method increase an Average Precision of 0.1%-1.3%. On Crowd Face, it increases an Average Precision of 1% - 6%. Dataset is available on: https://github.com/QxGeng/Crowd-Face.**

I. INTRODUCTION

Face detection is the key component of dealing with various face-related applications. In recent years, thanks to the development of Convolutional Neural Networks (CNNs), face detection has achieved great success [43], [42], [9]. Recently, many detectors based on deep learning have surpassed humans in visual related competitions. However, humans have a great advantage in dealing with low-resolution challenges because people can utilize rich domain knowledge flexibly [21]. Existing face detectors are challenging to handle small face detection in crowded scenes due to the presence of many small low-resolution faces and varying degrees of occlusion in crowded scenes.

Anchor-based face detectors have achieved satisfactory performance on the benchmark WIDER FACE [34]. Recently, many face detectors based on deep learning rely on features extracted from deep Convolutional Neural Network (CNN). They obtain low-level features of the objects such as texture information, edge information from the low layers of the network, and high-level features such as rich semantic information from the high layers of the network. However, for face detectors, thorny issues involved in detecting degraded faces are caused by small-size, defocus blur and occlusion in surveillance videos [37]. The appearance degradation caused by low resolution is the core problem of occluded and small face detection. These blur and low-resolution faces only have dozens or even a few pixels, so they contain limited feature information. When using the standard spatial pooling process [37] in a CNN, appearance features would be further degraded. CNN can only provide very few low-level features at the low layers, and there are almost no high-level features of these faces at the high layers. This problem is essentially ill-posed for a low-resolution ob-

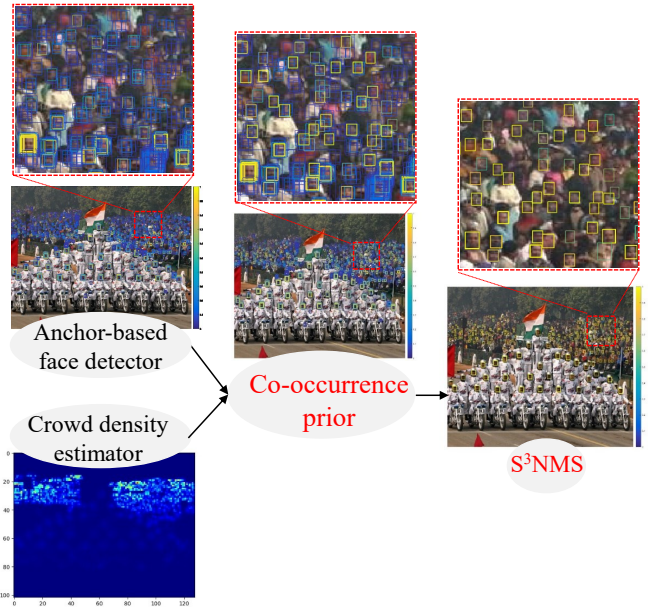

Fig. 1: The pipeline of our proposed post-processing approach. The prior of face co-occurrence can increase true positives of bounding-boxes according to crowd density estimation. S^{3}NMS further increases true positive and reduces false positive according to the inferred face boxes' score and size. The colorbar on the right of each image is the detector confidence, *i.e.*, blue boxes represent low confidence, and yellow boxes represent high confidence.

Dong Liang is the corresponding author

978-1-6654-3176-7/21/$31.00 ©2021 IEEE

ject. Therefore, a natural approach is to assemble information from the context.

In order to solve this problem, some works [8], [37], [41], [13] have introduced contextual information to make low-resolution faces contain more feature information. In anchor-based face detection methods, face contextual information is usually introduced in a low-level context, i.e., contextual information is obtained by modifying the size of the receptive field on the feature map. Obviously, rich low-level features are helpful to detect small objects, [4], [27] analyzes the role of contextual information for object detection in challenging situations. [2] shows that humans would make more errors in detecting objects that breach their standard context. In face detection, [37] focuses on detecting small faces with low-level context.

On the other hand, we argue that high-level contextual information is also valuable for small face detection, especially for the degenerate faces. Therefore, we explore the spatial contextual information and the relationship of objects as high-level contextual information. Unlike the low-level contextual information that adjusts the local receptive fields, our work utilizes contextual information on the entire image, not just around the object. Previous studies have proposed many video foreground segmentation methods based on high-level contextual information [11], [14], [15], [32], [31], [39], [33], [22], [12]. By this inspiration, in order to improve the utilization of spatial information, we introduce high-level contextual information to hard face detection.

In this paper, we propose a universal strategy with face co-occurrence prior based on density map and score-size-specific non-maximum suppression, independent of training paradigms to directly replace the standard non-maximum suppression (NMS) post-processing formula in anchor-based detectors. Specifically, we tap high-level spatial contextual information based on the crowd density map to detect the occurrence of degenerate faces, which we call co-occurrence prior. The prior of face co-occurrence coordinates outputs of the detector. It improves the detector's specificity and sensitivity by increasing true positives. We also propose score-size-specific non-maximum suppression for better removing redundant boxes in crowd scenes. It reduces false positives and increases true positives according to the inferred bounding-boxes' score and size. Fig. 1 shows the pipeline of our proposed approach. We can observe that, after integrate with our method, the detector can find more true faces. We also propose a dataset - Crowd Face for face detection, which is full of challenges for tiny faces. We expect to supply enough samples to highlight the difficulties of detecting dense and degenerate faces.

II. RELATED WORK

A. Face Detection

Face detection has derived benefit from the development of generic object detection [24], [18], [16], [17], [23]. The most advanced face detectors are built upon the anchor-based detection paradigm. S^3FD [36] indicates that multi-scale features perform better for tiny faces and it predicts boxes on multiple layers of feature hierarchy. [40] adopts a new anchor matching strategy to improve the recall rate of tiny faces. [1] introduces GAN-based super-resolution into face detection to compensate for the features of low-resolution faces. PyramidBox [28] takes full advantage of contextual information to provide additional supervision for tiny faces. DSFD [9] is an advanced face detector that constructs a pseudo-two-stage structure to achieve a fine-grained consideration of faces. ProgressFace [42] adopts a novel scale-aware progressive training mechanism to address large scale variations for face detection. TinaFace [43] indicates that methods presented in generic object detection can be used for handling tiny face detection and achieves state-of-the-art face detection performance. In this paper, we continue to tap the potential of anchor-based face detectors, expecting to enhance these methods' performance in crowd scenarios.

B. Context in Face Detection

Many works [21], [30], [4], [27], [37] have investigated the idea of utilizing context in object detection. For specific face detection algorithms, Hybrid Resolution Model (HR) [8] is an effective framework for finding tiny faces and it specifically shows that the large-scale receptive fields can be effectively encoded as a foveal descriptor that captures both coarse context and high-resolution image features. Similarly, [41] aggregates features around faces and bodies for detection, which greatly improves detection performance. The face contextual information is often introduced in the low-level context by acquiring different scales of the receptive field on the feature map. We expect to fit into a proper high-level context of a scene to improve the anchor-based face detectors.

C. Non-Maximum Suppression

The goal of Non-Maximum Suppression (NMS) [25] is to remove redundant inferred bounding-boxes, which has been an indispensable part of many anchor-based object detectors in computer vision in recent years [29], [7], [20], [26]. Soft-NMS [3] argues that traditional NMS is too greedy because only the highest scoring bounding-boxes are selected. Soft-NMS employs an approach to suppresses the bounding-boxes by reducing their score instead of removing it directly. More complex learning based post-processing methods rely on the model-related learning process. Hosang [6] proposed a learning-based NMS to improve localization and occlusion handling. Tychsen-Smith [29] argued that many detection methods are designed to identify only a sufficiently accurate bounding box, rather than the best available one, and proposed fitness NMS. We tend to develop a paradigm to better remove the redundant boxes, which is model-independent and can be embedded into existing face detectors without additional learning.

D. Density Map Based Crowd Counting

Crowd analysis often utilizes a density map because it not only exhibits the headcount, but also provides the location and spatial distribution of people. [38] proposes geometry

adaptive and fixed kernels with Gaussian convolution to generate a density map. [10] introduces a dilated convolutional neural network to improve the density map's quality. CAN [19] introduces an end-to-end trainable deep architecture that combines features obtained using multiple receptive field sizes and learns the importance of features at each image location, which adaptively encodes the scale of the contextual information needed to predict crowd density accurately. In this paper, crowd density estimation is employed to derive face co-occurrence prior for harmonizing a face detector's outputs.

III. THE PROPOSED APPROACHES

A. The Prior of Face Co-occurrence Based on Density Map

1) Crowd density map: Given a set of N training images $\{I_i\}_{(1\leq i\leq N)}$ and corresponding ground-truth density maps D_i^{gt}, the goal of density map estimation is to learn a non-linear mapping $\mathcal{F}$ that maps the input image I_i to an estimated density map $D_i^{est}(I_i) = \mathcal{F}(I_i)$, that approximates the ground truth D_i^{gt} in the L_2 norm criterion. To represent the density maps, for the image I_i, we define an array of 2-dimensional points $P_i^{est} = \{P_{i,j}\}_{1\leq j\leq C_i}$, which represents the location of each head in the scene, where C_i denotes the number of head in the image I_i. We convolve the image with the Gaussian kernel which denoted as $\mathcal{N}^{est}\left(p \mid \mu, \sigma^2\right)$, and then we obtain the estimated density map D_i^{est} by the total probability formula, as follows.

$$D_i^{est}\left(p \mid I_i\right) = \mathcal{F}(p \mid I_i) = \mathcal{F}[\sum_{j=1}^{C_i} \mathcal{N}^{gt}\left(p \mid \mu = P_{i,j}, \sigma^2\right)], \tag{1}$$

where μ denotes the mean of the normal distribution, and σ denotes the standard deviation.

In each input image, for head point $P_{i,j}$, we use $\{d_k^{i,j}\}_{(1\leq k\leq K)}$ to denote the distances to its K nearest neighbors. We utilize the following formula to represent the average distance between heads.

$$\overline{d^{i,j}} = \frac{1}{K}\sum_{k=1}^{K} d_k^{i,j}. \tag{2}$$

The size of the human head cannot be directly shown by estimated crowd density map. However, due to the dense distribution of individuals in a high-density crowded scene, the head size can be roughly represented by the distance. Thus, in crowded scenes, the distance between the centers of two neighboring individual is approximately equal to the head size. We utilized Context-Aware Network (CAN) [19] to estimate the density. It adaptively encodes the scale of context information by combining the features obtained from multiple receptive fields. Thus, it can estimate an accurate density map.

2) Face Co-occurrence: In this part, we focus on using the prior of face co-occurrence which based on density map to optimize the detectors in crowd scenarios. In a crowd scene, since the face size approaches the limit of imaging resolution, the appearance of face is inadequate and scarce.

Algorithm 1: The prior of face co-occurrence for inferred bounding-boxes coordination

Data: $\mathcal{B} = \{b_{x,y}\}$, $\mathcal{S} = \{s_{x,y}\}$, $\mathcal{A} = \{A_i^n\}$, D_i^{est}, γ, s_t,

B denotes the array of inferred bounding-boxes, S denotes the scores of bounding-boxes, A_i^n denotes the array of different density regions, D_i^{est} denotes the estimated density map.

for $b_{x,y}$ *in* $\mathcal{B}$ **do**
- $BS_{x,y} \leftarrow size(b_{x,y})$
- **for** $\mathcal{B}$ *in* A_i^n **do**
  - $a_i^n \leftarrow size(A_i^n)$; $\widehat{Z}_i^n \leftarrow \sum_{p\in A_i^n} D_{est}(p|A_i^n)$;
  - $\rho_i^n = \widehat{Z}_i^n / a_i^n$
  - **if** $s_{x,y} \geqslant s_t$ **then**
    - $m \leftarrow m+1$; $BS_{sum}^n \leftarrow BS_{sum}^n + BS_i$
  - $BS_{avg}^n \leftarrow BS_{sum}^n / m$

for $b_{x,y}$ *in* $\mathcal{B}$ **do**
- **if** $b_{x,y}$ *in* A_i^n **then**
  - **if** $(1-\gamma)BS_{avg}^n \leq BS_{x,y} \leq (1+\gamma)BS_{avg}^n$ **then**
    - $s_{x,y} = \sigma[D_{i(x,y)}^{est}(p \mid b_{x,y})\rho_i^n]s_{x,y} + s_{x,y}$

So, face detector would assign low confidence scores to these low-resolution faces and then remove these bounding-boxes by score threshold. Therefore, face co-occurrence is utilized to perform more responsive detection when faces are blurred or barely visible in crowd scenes. Specificallyin the high-density crowd scene, the distance between two neighboring heads is approximately equal to the head size, as mentioned earlier. Thus, we can find that the face is roughly the same size in the local area around each head. Therefore, a reasonable inference is that if many faces' score dominate in a region, then some inferred bounding-boxes of similar size to these faces are likely to be real faces. We increase these faces' scores according to the prior of face co-occurrence after the detector outputs the detection results.

So, we need to design a scheme to intervene and coordinate face detection in high-density regions which full of small and low-resolution faces. As mentioned before, the density map uses the pixel intensity of the map to show the distribution of the head. Thus, as shown in Algorithm 1, we proposed a density-map-based co-occurrence face detection strategy. Firstly, we utilize the density estimate network to obtain the density estimation map D_i^{est} of the input image. We generate blocks $\mathcal{A} = \{A_i^n\}$ on the image I_i by sliding windows, and to minimize the boundary effects, we adopt an overlap of 50%, where n represents the blocks' number. We integral over the values of the the predicted density map to estimate the number of people in each block, as follows,

$$\widehat{Z}_i^n = \sum_{p\in A_i^n} D_i^{est}(p \mid A_i^n). \tag{3}$$

Then we utilize the following formula to calculate the density

of each block and record it as ρ_i^n.

$$\rho_i^n = \widehat{Z}_i^n / a_i^n, \tag{4}$$

where a_i^n represents the area of region A_i^n. We designed two constraints to select the inferred bounding-boxes for coordination. (1) In each high-density block, an inferred bounding-box $s_{x,y}$ may be a real face if the score of the inferred bounding-box $s_{x,y}$ surpasses the threshold of score s_t. Then, we calculate all high score faces' average size and record it as BS_{avg}^n to tell us faces' size which appear in this region. (2) We further filter out bounding-boxes with area between $(1-\gamma, 1+\gamma)BS_{avg}^n$ as coordinated inferred bounding-boxes, and the boxes with final scores less than the score threshold will be removed. The coordinated formula is as follows,

$$s_{x,y} = \sigma[D_{i(x,y)}^{est}(p \mid b_{x,y})\rho_i^n]s_{x,y} + s_{x,y}, \tag{5}$$

where σ denotes the Sigmoid function, $b_{x,y}$ denotes the inferred bounding-box and $s_{x,y}$ denotes the corresponding confidence score. In this way, we can detect more real faces which could be ignored.

B. Score-size-specific NMS

NMS [25] is utilized as standard post-processing for object detection to partition bounding-boxes into non-overlapping subsets. The final detections are obtained by averaging the coordinates of the detection boxes in set B. If b_u and b_v are two bounding boxes, IoU refers to the standard Jaccard similarity (intersection over union overlap, IoU) used in NMS, which can be expressed as follows,

$$IoU(b_u, b_v) = \frac{b_u \cap b_v}{b_u \cup b_v}. \tag{6}$$

The traditional NMS retains the inferred bounding-box with the highest score and discards all the other inferred bounding-boxes which overlapped with an IoU threshold. Specifically, if $IoU(b_u, b_v) > N_t$, then the bounding-box with the lower score will be removed directly. NMS tends to ensure that the same face only have one bounding box. However, this approach will result in missed detection, and the face that is partially covered by another face cannot be detected.

To deal with this problem, Soft-NMS [3] provides a chance to preserve the overlapped objects using a function of penalizing the inferred scores. It decays the detection scores of all other objects with a continuous penalty function which has no penalty when there is no overlap and a large penalty at a high overlap. Soft-NMS updates the pruning step with a Gaussian penalty function as follows,

$$s_{x,y} = s_{x,y} e^{-\mathrm{IoU}(b_u, b_v)^2/\delta}. \tag{7}$$

This update rule is applied in each iteration, and scores of all the remaining detection boxes are updated. It suppresses the inferred box by decreasing its score rather than just removing it. Finally, the bounding-box will be deleted if its score is lower than the score threshold. However, in our early experiments, we observed that Soft-NMS can cause the increase of false positives because some redundant boxes cannot be removed due to their final penalized scores still higher than the threshold. Therefore, we need to make careful consideration of scores of the bounding-boxes to better remove redundant boxes.

Algorithm 2: Score-size-specific NMS

Data: $\mathcal{B} = \{b_{x,y}\}$, $\mathcal{S} = \{s_{x,y}\}$, N_t, S_t, B_t
$\mathcal{B}$ denotes the array of inferred bounding-boxes, $\mathcal{S}$ denotes the scores of bounding-boxes, N_t denotes the IoU threshold, S_t denotes the score threshold, B_t denotes the ACB threshold.

```
for b_{x,y} in B do
    b_m ← argmax(S)
    if IoU(b_m, b_{x,y}) ⩾ N_t or ACB(b_m, b_{x,y}) ⩾ B_t then
        if s_{x,y} ⩾ S_t then
            s_{x,y} = s_{x,y} e^{-IoU(b_u,b_v)^2/δ}
        else
            B ← B − b_{x,y}; S ← S − s_{x,y}
```

These two methods also ignore the role of boxes' size in the inferred boxes aggregation. Considering the most extreme situation that the areas of the two boxes are quite different, the b_u is very big, and b_v is very small, from the definition of formula (6), the intersection is much smaller than the union, and the $IoU(b_u, b_v)$ lower than the threshold of removing redundant boxes in NMS and Soft-NMS. In the inferred box aggregation process, we need to comprehensively consider the score and size of bounding-boxes to design a more reasonable method to implement removal and retention operations. Based on IoU, we propose ACB (Area Consistency of boxes), which is defined as follows,

$$ACB(b_u, b_v) = \frac{b_u \cap b_v}{min(b_u, b_v)}. \tag{8}$$

We adopt a constraint that if $ACB(b_u, b_v)$ higher than B_t (the value we choose is 0.9), the box with a lower score will be considered as a redundant box. Algorithm 2 shows our proposed algorithm. If $IoU(b_m, b_{x,y}) \geqslant N_t$ or $ACB(b_m, b_{x,y}) \geqslant B_t$, where b_m is the box with the higher score in $\mathcal{B}$, it decays the scores using a continuous function $s_{x,y} = s_{x,y} e^{-\mathrm{IoU}(b_u, b_v)^2/\delta}$. The box with high score is more likely to be an occluded face, and Soft-NMS is used to re-identify such a case. For a box with low score, NMS avoids this non-face box to be false positive.

Score-size-specific NMS is a compromise solution of NMS and Soft-NMS, which provides a fine-grained consideration of the score and the size to avoid arbitrary discarding or preservation of the bounding-boxes, which is essential in the task of multi-scale face detection. More detailed performance evaluation will be discussed in the experiment section.

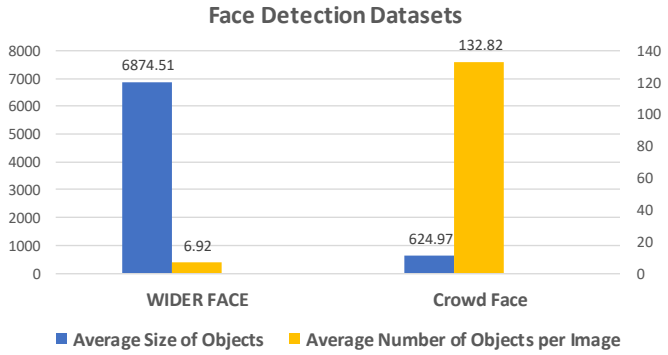

Fig. 2: Comparison of our proposed Crowd Face dataset and WIDER FACE dataset.

IV. EXPERIMENTAL EVALUATION

A. Dataset Preparation and Experimental Setting

WIDER FACE [34] is a benchmark dataset that is widely used in face detection studies. It has a total of 32203 images, of which 33793 faces are labeled. $4/10$ of the original images are randomly selected to form the training set, $1/10$ as the validation set, and $5/10$ as the testing set. The test set was divided into three subsets, according to the detection difficulty: "easy", "medium", and "hard", gradually increases various difficult situations in various face detection scenes in open environments.

Our proposed approach focuses on low-resolution and blurred face detection in crowd scenes, so we propose a challenging dataset - Crowd Face. It has a total of 34 images, of which 10371 faces are labeled, and the largest number of faces on a single image is 1101. As shown in Fig. 2, compared to WIDER FACE, each image in Crowd Face has more and smaller faces. As illustrated in Fig. 5, images from Crowd Face dataset have many low-resolution, small, and obscured faces. It is a challenging dataset with difficult samples, specifically for high-density face detection. Testing face detection algorithms on Crowd Face is helpful to explore the shortages of face detectors.

Experimental Setting In our experiments, we validate our proposed approach with the following state-of-the-art model, Hybrid Resolution Model (HR) [8], Single Shot Scale-invariant Face detector (S^3FD) [36], Light and Fast Face Detector (LFFD) [5], Context-anchor Hybrid Resolution Model (CAHR) [31], PyramidBox [28], Dual Shot Face Detector (DSFD) [9], Extremely Tiny Face Detector (EXTD) and TinaFace [43]. For the hyperparameters in our approach, we finally experimentally set $s_t = 0.5$, $\gamma = 0.1$ for face co-occurrence prior, $N_t = 0.3$, $S_t = 0.5$, $B_t = 0.9$ for score-size-specific NMS. Our experiments are run on GTX1080 with 16 GB RAM and 12-core i7 CPU.

B. Experiments for The Prior of Face Co-occurrence

In this part, in order to verify the performance of our proposed face co-occurrence prior based on density map in crowd scenes, we validate it on Crowd Face dataset.

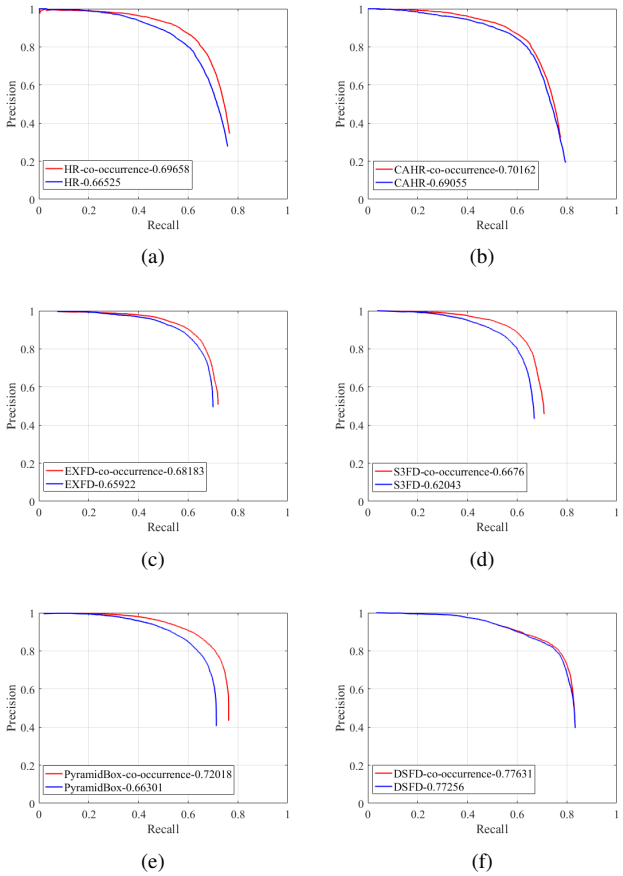

Fig. 3: Comparison between our proposed face co-occurrence and the original models on P-R curves.

We introduce density information on the basis of the state-of-the-art anchor-based detectors, and then combine with our proposed algorithm. As is shown in Fig. 3, compared with original detectors, the results show that our approach has higher Average Precision (AP) performance around 0.3-5.7%. Compared with the original detectors, all the advanced detectors have improved the performance after embedding our approach, indicates the capability of the proposed approach in challenging situations.

C. Experiments for Score-size-specific NMS

TABLE I: Average Precision (AP) performance of NMS, Soft-NMS and our proposed S^3NMS for HR, CAHR and PyramidBox on WIDER FACE hard and Crowd Face sets.

Data/Method	NMS	Soft-NMS	S^3NMS	Tested model
WIDER FACE hard	0.816	0.820	**0.827**	HR
	0.832	0.835	**0.843**	CAHR
	0.888	0.889	**0.890**	PyramidBox
Crowd Face	0.665	0.683	**0.707**	HR
	0.691	0.707	**0.720**	CAHR
	0.663	0.671	**0.681**	PyramidBox

In this part, we test our proposed score-size-specific NMS on the WIDER FACE hard set and Crowd Face set. Our

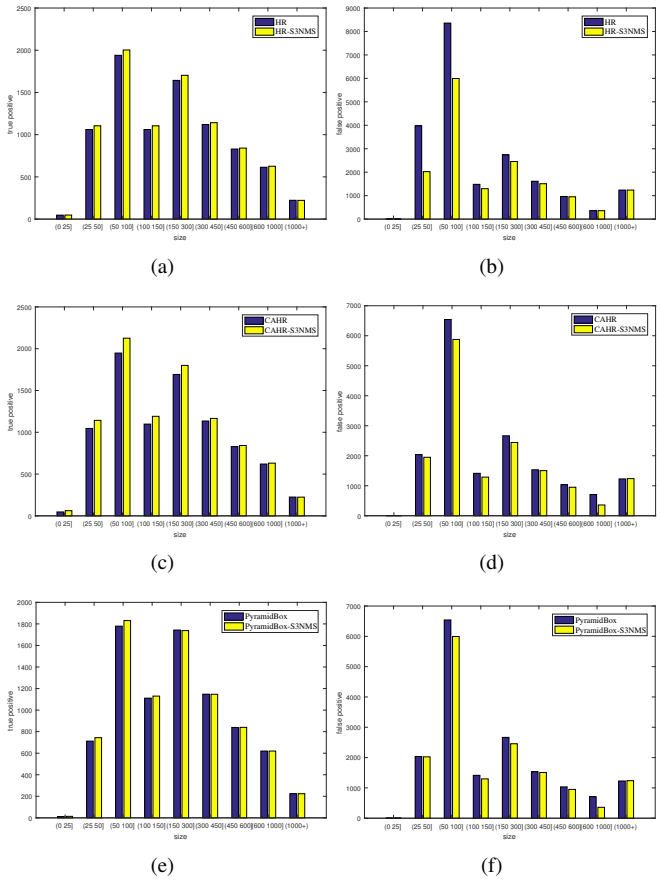

Fig. 4: Comparison of true and false positives for original HR, CAHR, PyramidBox and these models with our proposed score-size-specific NMS.

proposed S^{3}NMS is a post-processing method without any additional training. We compared our approach with other post-processing methods NMS and Soft-NMS, which also do not need to re-train the model. We respectively integrate NMS, soft-NMS, and our proposed S^{3}NMS into anchor-based detectors including HR [8], CAHR [31] and PyramidBox [28]. As shown in Table I, S^{3}NMS has the highest Average Precision (AP) compared with NMS and soft-NMS on WIDER FACE hard set and Crowd Face set. It illustrates that we need a fine-grained consideration of the score and the size to remove redundant boxes. Fig. 4 shows the comparison of true and false positives for baseline models and these models integrated with our proposed S^{3}NMS on Crowd Face. It illustrates that our method can reduce false positives and increase true positives in crowd scenes.

D. Ablation Study on Crowd Face

As shown in Table II, we perform ablation experiments on Crowd Face. We separately integrate NMS, score-size-specific NMS, and co-occurrence prior to HR, PyramidBox, EXTD, CAHR, DSFD and TinaFace on the Crowd Face set. We first compare the performance of NMS and our proposed S^{3}NMS, it shows that our proposed S^{3}NMS has higher

TABLE II: Ablation study of our proposed co-occurrence prior based on density map and score-size-specific NMS integrated with HR, PyramidBox, EXTD, CAHR, DSFD and TinaFace on Crowd Face.

Method	NMS	S^{3}NMS	Co-occurrence.	$AP(\%)$
HR [8]	✓			0.665
		✓		0.707
	✓		✓	0.697
		✓	✓	**0.710**
PyramidBox [28]	✓			0.663
		✓		0.681
	✓		✓	0.720
		✓	✓	**0.725**
EXTD [35]	✓			0.659
		✓		0.674
	✓		✓	0.682
		✓	✓	**0.688**
CAHR [31]	✓			0.691
		✓		0.720
	✓		✓	0.702
		✓	✓	**0.728**
DSFD [9]	✓			0.772
		✓		0.780
	✓		✓	0.776
		✓	✓	**0.781**
TinaFace [43]	✓			0.771
		✓		0.776
	✓		✓	0.781
		✓	✓	**0.784**

AP performance. Then, we respectively integrate NMS and S^{3}NMS with our co-occurrence prior into the detectors. The result shows our proposed co-occurrence prior can further improve the performance, and S^{3}NMS combined with co-occurrence prior has the best AP performance. Our results increase an AP of 1% - 6%. Fig. 5 shows Some visualized comparison of the original HR detector (magenta rectangles) and embed our proposed approach to HR (cyan ellipses) in crowd scenes. The proposed method achieves notably better precision as it can detect more true faces. It illustrates that the proposed method can enhance the detectors to find more true faces in crowd scenes when there are many low-resolution small faces.

E. Overall Performance on WIDER FACE

We verify our proposed co-occurrence prior and score-size-specific NMS on the WIDER FACE dataset in this part. As illustrated in Table III, we embed our propose approach to some face detectors and compare the AP performance between the original detectors and our approach on the WIDER FACE dataset. The results show that embedding our proposed method into all face detectors in WIDER FACE-hard set have better performance than the original detectors, indicating the capability of the proposed approach in challenging situations. Especially the performance improvement is obvious by embedding our proposed method to detectors with poor inferring performance. As WIDER FACE-easy set contains almost no high-density scenes, our face co-occurrence based on density information cannot find more faces, but our method does not deteriorate the original performance in low-density scenarios.

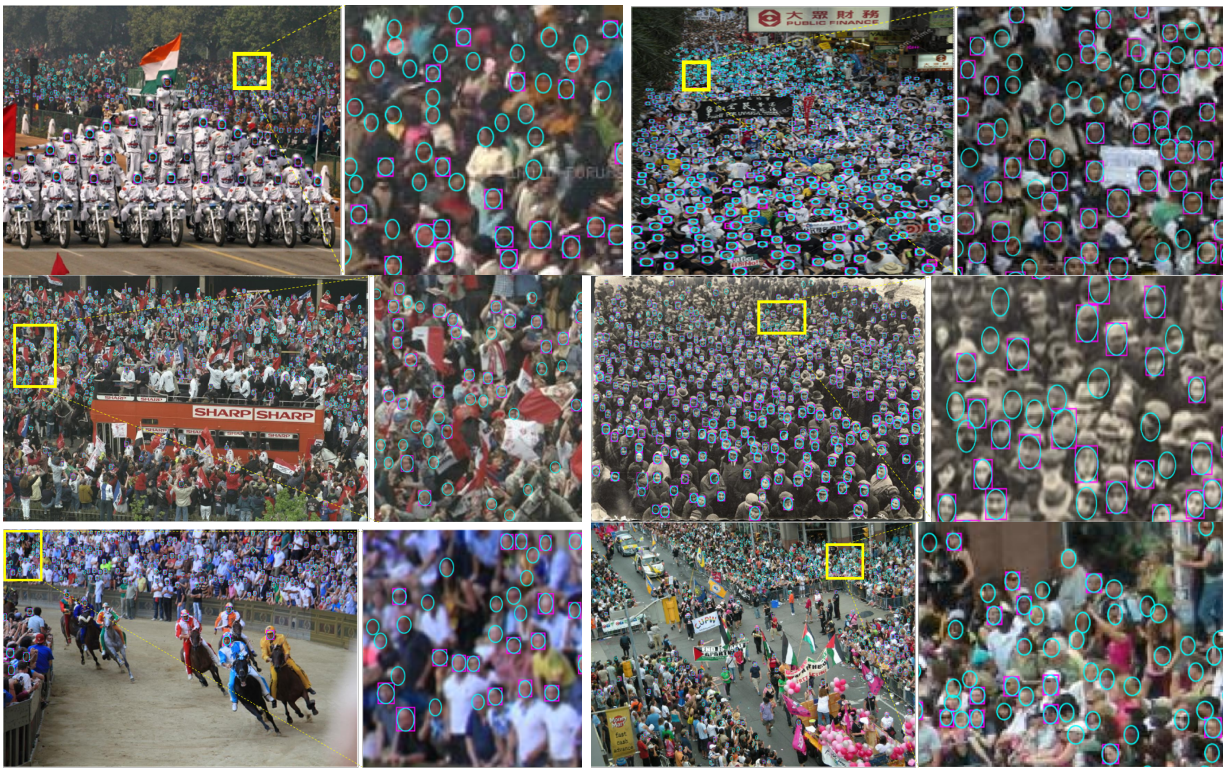

Fig. 5: Some visualized comparison of the original HR detector (magenta rectangles) and embed our proposed approach to HR (cyan ellipses) in crowd scenes.

TABLE III: Performance of integrating co-occurrence prior and score-size-specific NMS to the trained detectors on WIDER FACE.

Sub-set in WIDER FACE	easy		medium		hard	
Method	Orignal	Proposed	Orignal	Proposed	Orignal	Proposed
LFFD [5]	0.873	**0.876**	0.861	**0.865**	0.750	**0.758**
HR [8]	0.925	0.925	0.911	**0.912**	0.816	**0.829**
CAHR [31]	0.928	0.928	0.912	**0.913**	0.832	**0.844**
EXTD [35]	0.921	**0.923**	0.911	**0.912**	0.846	**0.853**
S^3FD [36]	0.945	0.945	0.934	**0.936**	0.853	**0.855**
PyramidBox [28]	0.960	0.960	0.948	**0.950**	0.888	**0.890**
DSFD [9]	0.966	0.966	0.957	0.957	0.905	**0.906**
TinaFace [43]	0.963	**0.964**	0.956	**0.958**	0.930	**0.932**

V. CONCLUSION

In this paper, we propose a general approach with density-map-based face co-occurrence prior by mining high-level spatial contextual information, and score-size-specific non-maximum suppression (S^3NMS) according to the inferred face boxes' score and size. The co-occurrence prior can detect more real faces in crowded scenes, which is important for us to break the challenge of low-resolution and small faces in crowded scenes. S^3NMS avoids arbitrary discarding or preservation of the bounding box and reduces false positives and increases true positives. Our proposed approach does not need any additional learning and is easy to implement. In the future, we will further focus on the challenges in crowd scenes.

REFERENCES

[1] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem. Finding tiny faces in the wild with generative adversarial network. In *CVPR*, pages 21–30, 2018.
[2] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz. Scene perception" detecting and judging objects undergoing relational violations. In *Cognitive Psychology*, pages 143–177, 1982.
[3] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis. Soft-nms improving object detection with one line of code. In *ICCV*, pages 5562–5570, 2017.
[4] S. K. Divvala, D. W. Hoiem, J. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *CVPR*, pages 1271–1278, 2009.
[5] Y. He, D. Xu, and L. Wu. Lffd: A light and fast face detector for edge devices. In *arXiv*, 2019.
[6] J. Hosang, R. Benenson, and B. Schiele. Learning non-maximum suppression. In *CVPR*, pages 4507–4515, 2017.
[7] J. H. Hosang, R. Benenson, and B. Schiele. A convnet for non-maximum suppression. In *GCPR*, pages 192–204, 2016.
[8] P. Hu and D. Ramanan. Finding tiny faces. pages 1522–1530, 2017.
[9] J. Li, Y. Wang, and C. Wang. Dsfd: Dual shot face detector. In *CVPR*, pages 5060–5069, 2019.

[10] Y. Li, X. Zhang, and D. Chen. Csr-net: Dilated convolutional neural networks for understanding the highly congested scenes. In *CVPR*, pages 1091–1100, 2018.
[11] D. Liang, M. Hashimoto, K. Iwata, X. Zhao, et al. Co-occurrence probability-based pixel pairs background model for robust object detection in dynamic scenes. *Pattern Recognition*, pages 1374–1390, 2015.
[12] D. Liang, S. Kaneko, and Y. Satoh. A robust appearance model and similarity measure for image matching. *Journal of Robotics and Mechatronics*, pages 126–135, 2015.
[13] D. Liang, S. Kaneko, H. Sun, and B. Kang. Adaptive local spatial modeling for online change detection under abrupt dynamic background. In *ICIP*, pages 2020–2024, 2017.
[14] D. Liang, B. Kang, X. Liu, H. Sun, L. Zhang, and N. Liu. Cross scene video foreground segmentation via co-occurrence probability oriented supervised and unsupervised model interaction. In *ICASSP*, pages 1795–1799, 2021.
[15] D. Liang and X. Liu. Coarse-to-fine foreground segmentation based on co-occurrence pixel-block and spatio-temporal attention model. In *ICPR*, pages 3807–3813, 2021.
[16] T. Lin, P. Dollr, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 936–944, 2017.
[17] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollr. Focal loss for dense object detection. In *TPAMI*, pages 318–327, 2017.
[18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *ECCV*, pages 21–37, 2016.
[19] W. Liu, M. Salzmann, and P. Fua. Context-aware crowd counting. In *CVPR*, pages 5099–5108, 2019.
[20] A. Neubeck and L. Van Gool. Efficient non-maximum suppression. In *ICPR*, pages 850–855, 2006.
[21] A. Oliva and A. Torralba. The role of context in object recognition. In *Trends in Cognitive Sciences*, pages 520–527, 2007.
[22] J. Pan and D. Liang. Holistic crowd interaction modelling for anomaly detection. In *CCBR*, pages 642–649, 2017.
[23] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. In *arXiv*, 2018.
[24] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *TPAMI*, pages 1137–1149, 2017.
[25] A. Rosenfeld and M. Thurston. Edge and curve detection for visual scene analysis. In *IEEE Transactions on Computers*, pages 562–569, 1971.
[26] R. Rothe, M. Guillaumin, and L. Van Gool. Non-maximum suppression for object detection by passing messages between windows. In *ACCV*, pages 290–306, 2014.
[27] A. Shrivastava, A. Gupta, and R. B. Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, pages 761–769, 2016.
[28] X. Tang, D. K. Du, and Z. He. Pyramidbox: A context-assisted single shot face detector. In *ECCV*, pages 797–813, 2018.
[29] L. Tychsensmith and L. Petersson. Improving object localization with fitness nms and bounded iou loss. In *CVPR*, pages 6877–6885, 2018.
[30] L. Wolf and S. M. Bileschi. A critical view of context. In *IJCV*, pages 251–261, 2006.
[31] T. Wu, D. Liang, J. Pan, and S. Kaneko. Context-anchors for hybrid resolution face detection. In *ICIP*, pages 3297–3301, 2019.
[32] T. Wu, D. Liang, J. Pan, H. Sun, B. Kang, S. Kaneko, and H. Zhou. Score-specific non-maximum suppression and coexistence prior for multi-scale face detection. In *ICASSP*, pages 1957–1961, 2019.
[33] S. Xiang, D. Liang, S. Kaneko, and H. Asano. Robust defect detection in 2d images printed on 3d micro-textured surfaces by multiple paired pixel consistency in orientation codes. *IET Image Processing*, pages 3373–3384, 2020.
[34] S. Yang, P. Luo, C. C. Loy, and X. Tang. Wider face: A face detection benchmark. In *CVPR*, pages 5525–5533, 2016.
[35] Y. Yoo, D. Han, and S. Yun. Extd: Extremely tiny face detector via iterative filter reuse. In *arXiv*, 2019.
[36] S. Zhang, X. Zhu, and Z. Lei. S^3fd: Single shot scale-invariant face detector. In *ICCV*, pages 192–201, 2017.
[37] S. Zhang, X. Zhu, Z. Lei, X. Wang, H. Shi, and S. Z. Li. Detecting face with densely connected face proposal network. In *CCBR*, pages 3–12, 2018.
[38] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, pages 589–597, 2016.
[39] W. Zhou, S. Kaneko, M. Hashimoto, Y. Satoh, and D. Liang. Foreground detection based on co-occurrence background model with hypothesis on degradation modification in dynamic scenes. *Signal Processing*, pages 66–79, 2019.
[40] C. Zhu, R. Tao, K. Luu, and M. Savvides. Seeing small faces from robust anchor's perspective. In *CVPR*, pages 5127–5136, 2018.
[41] C. Zhu, Y. Zheng, K. Luu, and M. Savvides. Cms-rcnn: Contextual multi-scale region-based cnn for unconstrained face detection. In *Deep Learning for Biometrics*, pages 57–79, 2017.
[42] J. Zhu, D. Li, T. Han, L. Tian, and Y. Shan. Progressface: Scale-aware progressive learning for face detection. In *ECCV*, pages 344–360, 2020.
[43] Y. Zhu, H. Cai, S. Zhang, C. Wang, and Y. Xiong. Tinaface: Strong but simple baseline for face detection. In *arXiv*, 2020.