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

***Abstract*— Though object detection has made tremendous strides, small face detection remains one of the key challenges in face detection. A central issue of small face detection is the appearance degradation caused by shallow resolution. Therefore, aggregating information from context becomes a natural choice. This paper discusses how to properly utilize high-level contextual prior to enhance the capabilities of anchor-based detectors for dense and degenerate face detection. We mine the spatial contextual information on the holistic view according to the density estimation and propose face co-occurrence prior for inferred box harmonization. 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 approach is plug and play and model-independent, which could be integrated into the existing anchor-based face detectors without extra learning. We also collect a challenging face detection dataset - Crowd Face, to provide adequate samples to prominent the bottleneck of detecting crowded faces. We integrate our proposed methods with state-of-the-art anchor-based face detectors on massively benchmarked face datasets (WIDER FACE and Crowd Face). When compared to 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

Robust face detection in open world is an ultimate component to handle various facial related problems. Due to the promising development of deep Convolutional Neural Networks, face detection has made tremendous progress in recent years [44], [43], [9]. Renewed detection paradigms [16], [8], strong backbone [18], [24], [17] and large scale dataset [35] jointly push forward the limit of face detection to approach humans' cognition that many detectors have surpassed humans on visual detection and recognition competitions. However, because flexible mechanisms and abundant domain knowledge guide human's cognition, human has advantages on handing the challenges of low resolution [21]. In video surveillance, since the faces are usually far away from the surveillance camera and have varying degrees of occlusion problems, small face detection in crowded scenes is a challenging problem with practical needs.

Dong Liang is the corresponding author

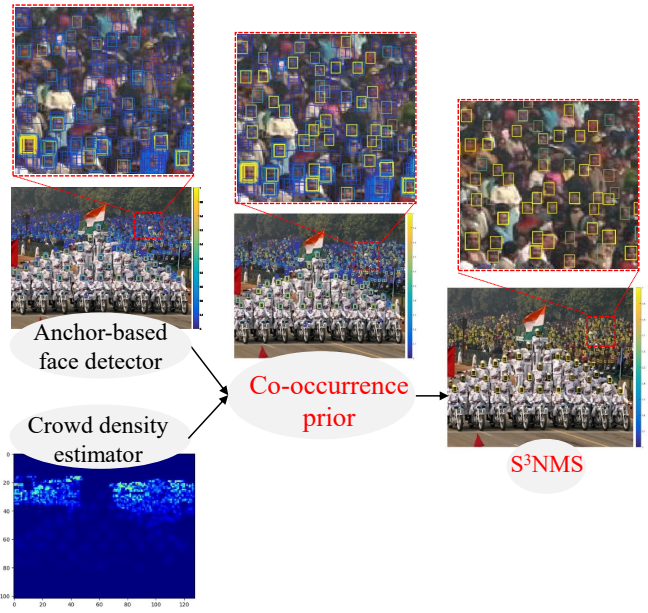

Fig. 1: Architecture of the proposed post-processing framework. Face co-occurrence prior increase true positives of the inferred face boxes according to crowd density estimation. S^{3}NMS further increases true positive and reduces false positive according to the inferred face boxes' score and size.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

Anchor-based face detectors have achieved satisfactory performance on the benchmark WIDER FACE [35]. 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 [38]. The central issue of small and occluded face detection in crowded scenes is the appearance degradation caused by shallow resolution. 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 [38] 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 object. Therefore, aggregating information from context becomes a natural choice.

In order to solve this problem, some works [8], [38], [42], [13] have introduced contextual information to make low-resolution faces contain more feature information. In anchor-based face detection methods, the contextual information of faces is usually employed in the low-level context via a different receptive field of feature maps. Obviously, rich low-level features are helpful to detect small objects, [4], [27] reviewed contextual information and analyzed its role for challenging object detection in empirical evaluation. [2] shows that humans detecting objects that violate their standard context take longer and make more errors. For face detection, [38] focus on low-level context to detect small faces and [8] demonstrates that both contextual information and scale-variant representations are crucial.

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. Different from low-level contextual information which adjusts the local receptive fields, our work extends the contextual information to the whole image rather than just surrounding objects. In our previous researches, we proposed a series of background modeling methods based on high-level contextual information to obtain stable context information between pixels for video foreground segmentation [11], [14], [15], [33], [32], [40], [34], [22], [12]. Inspired by this, in this paper, we try to introduce the high-level contextual information to hard face detection to improve the utilization efficiency of scene spatial information.

In this paper, we propose a universal strategy with density-map-based face co-occurrence prior 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 mine the high-level spatial contextual information according to crowd density estimation to detect the occurrence of degenerate faces, which we call co-occurrence prior. Face co-occurrence prior harmonizes the outputs of a detector. It enhances the sensitivity and specificity of the detector via 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 face boxes' score and size. Fig. 1 illustrates the proposed detection framework. We can observe that, after integrate with our method, the detector can find more true faces. We also collect a challenging face detection dataset with tiny faces (*i.e.*, Crowd Face) to provide adequate samples to further prominent the bottleneck of detecting crowded faces.

II. Related Work

A. Face Detection

Face detection has derived benefit from the development of generic object detection [24], [16], [17], [23]. Most recent state-of-the-art face detectors are built upon the anchor-based detection paradigm. S^3FD [37] indicates that multi-scale features perform better for tiny faces and it predicts boxes on multiple layers of feature hierarchy. [41] adopts a new anchor matching strategy to improve the recall rate of tiny faces. [1] introduces the super-resolution based on GAN to face detection to make up the feature of low-resolution faces. PyramidBox [28] fully exploits the context information to provide extra supervision for small faces. DSFD [9] constructs pseudo two-stage structure based on a single-shot framework to make the face detector more effective and accurate. ProgressFace [43] adopts a novel scale-aware progressive training mechanism to address large scale variations for face detection. TinaFace [44] 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

The ideas of using context in object detection have been studied in many works [21], [31], [4], [27], [38]. For specific face detection algorithms, Hybrid Resolution Model (HR) [8] is a simple yet effective framework for finding small faces and it specifically shows that massively-large receptive fields can be effectively encoded as a foveal descriptor that captures both coarse context and high-resolution image features. Similarly, [42] pools ROI features around faces and bodies for detection, which significantly improves overall performance. The context information of faces is usually employed in the low-level context by acquiring different feature maps' receptive fields. We expect to fit into a proper high-level context of a scene to enhance the anchor-based face detectors.

C. Non-Maximum Suppression

The goal of Non-Maximum Suppression (NMS) [25] has a positive impact on performance measures that penalize false detections, which has been an integral part of many object detection algorithms in computer vision for almost 50 years [30], [7], [20], [26]. Soft-NMS [3] argues that the conventional NMS is too greedy because only the bounding-box with the maximum score is selected. Soft-NMS employs an approach that suppresses the bounding box by reducing its score instead of just removing it. 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 plug and play and model-independent paradigms to better remove the redundant boxes, which could be integrated into the existing anchor-based face detectors without extra learning.

D. Density Map Based Crowd Counting

A density map is widely used in crowd analysis since it can exhibit the headcount, locations and spatial distribution. [39] 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. [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 required 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. Face Co-occurrence Prior Based on Density map

1) Crowd density map: Given a set of N training images $\{I_i\}_{(1\leq i\leq N)}$ with 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 an input image I_i to an estimated density map $D_i^{est}(I_i) = \mathcal{F}(I_i)$, that is close to the ground truth D_i^{gt} in term of L_2 norm. To represent the density maps, to each image I_i, we associate a set of 2D points $P_i^{est} = \{P_{i,j}\}_{1\leq j\leq C_i}$ that denote the position of each human head in the scene, where C_i is the headcount in image I_i. The corresponding estimated density map D_i^{est} is obtained by a total probability formula via convolving an image with a Gaussian kernel $\mathcal{N}^{est}\left(p \mid \mu, \sigma^2\right)$. We have $\forall p \in I_i$,

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

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

For each head point $P_{i,j}$ in a given image, denoting the distances to its K nearest neighbors as $\{d_k^{i,j}\}_{(1\leq k\leq K)}$. The average distance is therefore

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

A crowd density map cannot directly show the size of the head. However, in a high-density crowd scene, since the individuals are densely distributed, the distance can roughly represent the head size. The head size is approximately equal to the distance between two neighboring individuals' centers in crowded scenes. The density estimate network we used is Context-Aware Network (CAN) [19]. It adaptively encodes the scale of context information by combining the features obtained from multiple receptive fields. Therefore, it can accurately estimate the crowd density map, especially when perspective effects are strong.

Algorithm 1: Face co-occurrence prior for inferred box harmonization

Data: $\mathcal{B} = \{b_{x,y}\}$, $\mathcal{S} = \{s_{x,y}\}$, $\mathcal{A} = \{A_i^n\}$, D_i^{est}, γ, s_t,
B is the list of initial inferred boxes, S contains corresponding inferred scores, A_i^n is the list of different density areas, D_i^{est} is the estimated density map.

```
for b_{x,y} in B do
    BS_{x,y} ← size(b_{x,y})
    for B in A_i^n do
        a_i^n ← size(A_i^n); Ẑ_i^n ← Σ_{p∈A_i^n} D_est(p|A_i^n);
        ρ_i^n = Ẑ_i^n / a_i^n
        if s_{x,y} ⩾ s_t then
            m ← m + 1; BS_sum^n ← BS_sum^n + BS_i
        BS_avg^n ← BS_sum^n / m
for b_{x,y} in B do
    if b_{x,y} in A_i^n then
        if (1 − γ)BS_avg^n ≤ BS_{x,y} ≤ (1 + γ)BS_avg^n then
            s_{x,y} = σ[D^est_{i(x,y)}(p | b_{x,y})ρ_i^n]s_{x,y} + s_{x,y}
```

2) Co-occurrence of homogeneous faces: In this part, we focus on using the face co-occurrence prior 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 face appearance is scarce and inadequate. So, many faces are assigned low confidence scores by the detector and then filtered by the score threshold, resulting in failure to detect them. Therefore, we utilize the co-occurrence of faces to make more sensitive detection when the face is ambiguously or marginally visible in a crowd scene. Face co-occurrence here refers to the co-occurrence of homogeneous faces. Specifically, as mentioned earlier, the head size in a high-density crowd scene is approximately equal to the distance between two neighboring heads. Therefore, we can observe that in the local region around each head, the size of the face is approximately the same size. So, if the scores of many faces dominate in a local region, it is reasonable that some inferred boxes which are similar to the sizes of these faces have a high probability of being faces. According to the co-occurrence prior, we increase the scores of real faces with low scores after a detector's inferring phase.

Hence, we need to design a mechanism to reconcile and intervene face detection in high-density regions (small face regions), and give up interventions for low-density regions. As mentioned earlier, the density map presents how heads distribute in terms of the pixel intensity of the map. So, we propose a co-occurrence face strategy based on density map, as illustrated in Algorithm 1. We send the image into the

density estimate network to get the density map D_i^{est} first. We define a dense grid on image I_i, and produce blocks $\mathcal{A} = \{A_i^n\}$ with 50% overlapping to minimize border effects, where n is the number of blocks. The number of people in different blocks is estimated by integrating over the values of the predicted density map as follows,

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

We calculate the density of each corresponding block and record it as ρ_i^n.

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

where a_i^n is the area of region A_i^n. There are two constraints to filter the inferred box for reconciliation. (1) In the corresponding high-density block, if the score of an inferred box $s_{x,y}$ exceeds the score threshold s_t, the inferred box could be a true human face. Then, the average size of all the high score faces is calculated and recorded as BS_{avg}^n, which tells us the size of faces that appear in the region. (2) These boxes with the size between $(1-\gamma, 1+\gamma)BS_{avg}^n$ are further filtered out as the inferred box for reconciliation and the inferred boxes whose scores are ultimately lower than the original threshold will be deleted. The reconciliation 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 is the Sigmoid function, $b_{x,y}$ is the inferred box and $s_{x,y}$ is the corresponding confidence score. In this way, we increase the scores of real faces which have low confidence scores.

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 conventional NMS preserves the detection box with the maximum score and discards all the other inferred boxes overlapped with an IoU threshold. Specifically, if $IoU(b_u, b_v) > N_t$, (0.3 is obtained here as most detectors using this value), then the box with the lower score is deleted directly. NMS tends to guarantee that the same face corresponds to only one bounding box. This principle is also useful for the multi-scale pyramid scheme, as one face may be detected in different layers of the pyramid. However, this way will cause missed detection, and the face covered by part of 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 reducing its score instead of just removing it. Finally, if the score of the bounding box is lower than the threshold of score, then this bounding box will be deleted. However, in our early experiments, we observed that Soft-NMS can cause the increase of false positives because some redundant boxes cannot be deleted 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.

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 $IoU(b_u, b_v)$ cannot reach the threshold of deleting 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 $s_{x,y} = s_{x,y} e^{-\mathrm{IoU}(b_u, b_v)^2/\delta}$. It uses NMS when the bounding box's score is low and uses Soft-NMS when the score is

Algorithm 2: Score-size-specific NMS

Data: $\mathcal{B} = \{b_{x,y}\}$, $\mathcal{S} = \{s_{x,y}\}$, N_t, S_t, B_t
$\mathcal{B}$ is the list of initial inferred boxes, $\mathcal{S}$ contains corresponding inferred scores, N_t is the IoU threshold, S_t is the score threshold, B_t is 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}
```

high. A high score box is more likely to be an occluded face, and Soft-NMS is used to re-identify such a case. For a low score box, NMS avoids this non-face box to be false positive. It gives a chance to detect faces covered by other faces without causing false positives as Soft-NMS does.

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 box, which is essential in the multi-scale face detection task. More detailed performance evaluation will be discussed in the experiment section.

IV. EXPERIMENTAL EVALUATION

A. Dataset Preparation and Experimental Setting

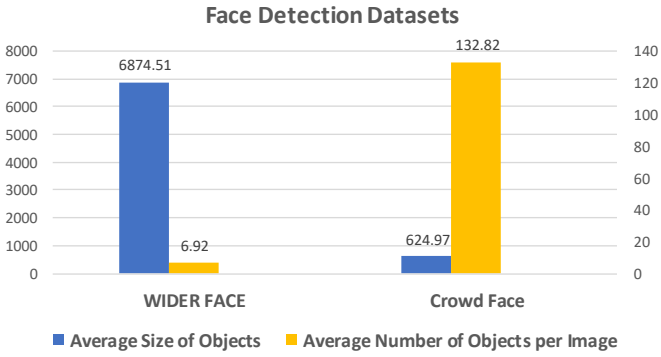

Fig. 2: Comparison of benchmark dataset WIDER FACE and our Crowd Face dataset. Two quantities are measured for each dataset: average size of objects (blue plots)and average number of objects per image (orange plots).

WIDER FACE In face detection literature, a widely used benchmark is WIDER FACE [35]. WIDER FACE contains 32203 images with 393793 faces, 40% of which are used for training, 10% for validation, and 50% for testing. According to the detection rate, the validation data are divided into three classes: "easy", "medium", and "hard", gradually increases various difficult situations in various face detection scenes in open environments.

Crowd Face Considering the proposed solution in this paper is mainly for low-resolution and obscured face detection in crowd scenes, we have prepared a harder dataset - Crowd Face. There are 34 images with 10731 annotated faces, and the maximum number of faces on an image is 1001. As illustrated in Fig. 2, each image in Crowd Face has smaller and more faces than WIDER FACE. As shown 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, the models we used to verify our proposed methods are , Hybrid Resolution Model (HR) [8], Single Shot Scale-invariant Face detector (S^{3}FD) [37], Light and Fast Face Detector (LFFD) [5], Context-anchor Hybrid Resolution Model (CAHR) [32],

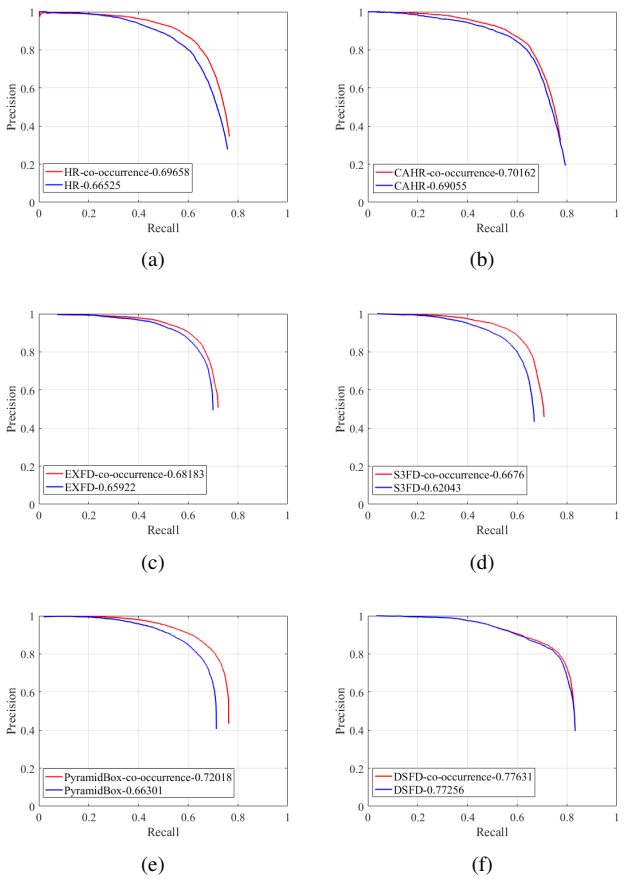

Fig. 3: P-R curve of face co-occurrence based on density maps, compared with the original models (HR, CAHR, EXTD, S^{3}FD, PyramidBox, DSFD).

PyramidBox [28], Dual Shot Face Detector (DSFD) [9], Extremely Tiny Face Detector (EXTD) and TinaFace [44]. All the models we used in the experiments are trained with the WIDER FACE training set and tested on the WIDER FACE validation set and Crowd Face. In our experiments, we compare many different settings of parameters, and finally 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 Face Co-occurrence Prior Based on Density map

In this part, in order to verify the performance of our proposed face co-occurrence prior based on density map in crowd scenarios, we test it on Crowd Face dataset. 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 illustrated in Fig. 3, Precision-Recall curves show that our method has higher Average Precision (AP) performance than original detectors around 0.3-5.7%. All the state-of-the-art anchor-based detectors enhance the performance after integrating our method compared with their

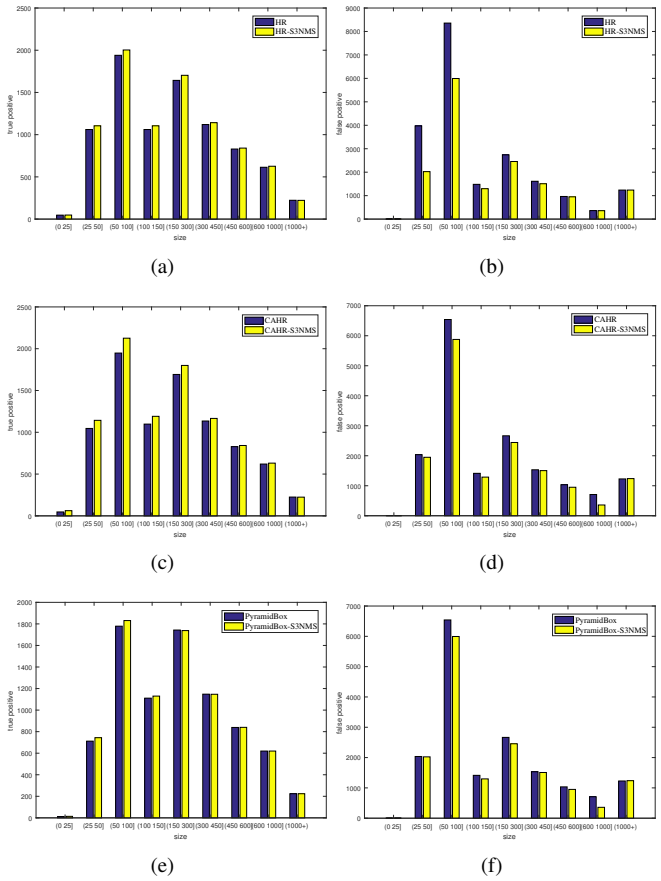

Fig. 4: Comparison of true and false positives for original HR, CAHR, PyramidBox and these models with our proposed score-size-specific NMS.

original detectors, 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^{3}NMS for HR, CAHR and PyramidBox on WIDER FACE hard and Crowd Face sets.

Data/Method	NMS	Soft-NMS	S^{3}NMS	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 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 [32] 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

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 [36]	✓			0.659
		✓		0.674
	✓		✓	0.682
		✓	✓	**0.688**
CAHR [32]	✓			0.691
		✓		0.720
	✓		✓	0.702
		✓	✓	**0.728**
DSFD [9]	✓			0.772
		✓		0.780
	✓		✓	0.776
		✓	✓	**0.781**
TinaFace [44]	✓			0.771
		✓		0.776
	✓		✓	0.781
		✓	✓	**0.784**

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 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 of the visual comparisons between our proposed method within HR (cyan ellipses) and original HR (magenta rectangles) 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.

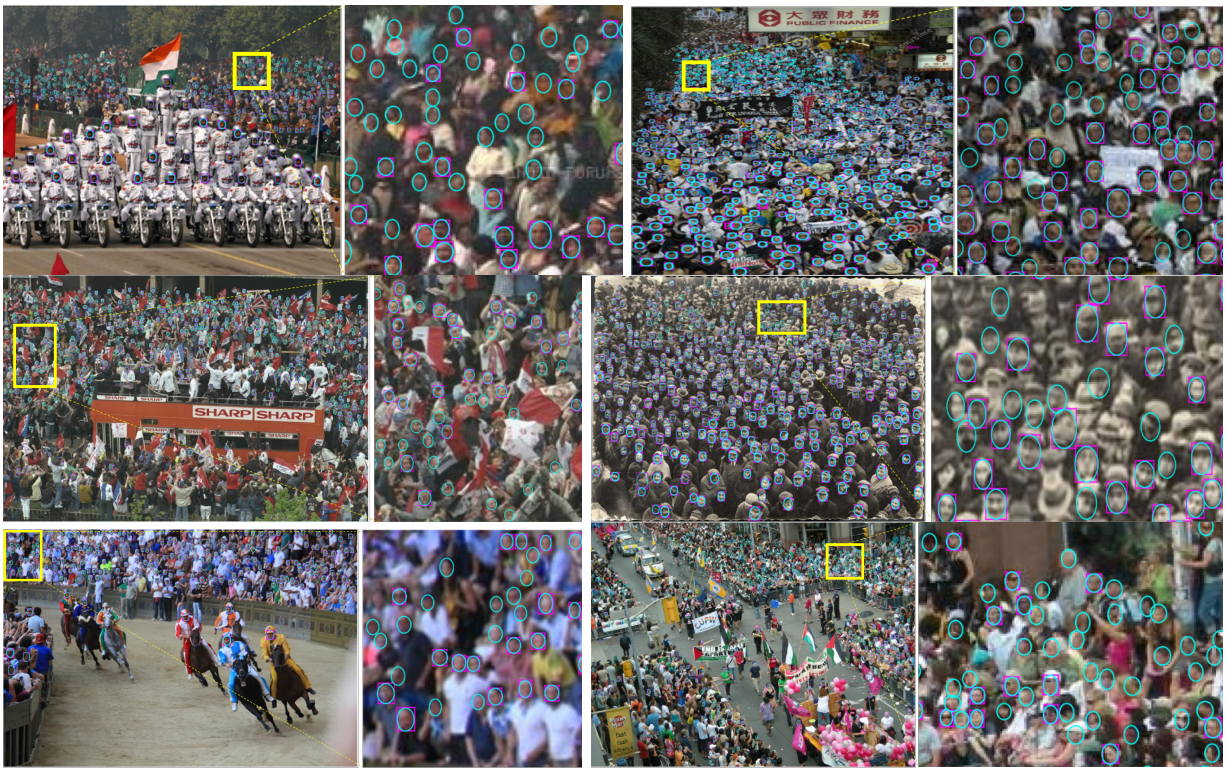

Fig. 5: Comparison of the proposed approach integrated into HR detector (cyan ellipses) and the original HR (magenta rectangles) in crowd scenes.

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 [32]	0.928	0.928	0.912	**0.913**	0.832	**0.844**
EXTD [36]	0.921	**0.923**	0.911	**0.912**	0.846	**0.853**
S^3FD [37]	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 [44]	0.963	**0.964**	0.956	**0.958**	0.930	**0.932**

E. Overall Performance on WIDER FACE

In this part, we test our proposed co-occurrence prior and score-size-specific NMS on the WIDER FACE dataset. As shown in Table III, we integrate our propose method to several anchor-based detectors including the state-of-the-art ones, and compare their AP performance with the original detectors on the WIDER FACE dataset. Table III shows that the proposed approach integrating within all face detectors have better performance than the original methods in WIDER FACE-hard set, indicating the capability of the proposed approach in challenging situations. Especially for the detectors with poor inferring performance, the improvements of the detectors integrated with the proposed scheme are obvious. As WIDER FACE-easy set contains almost no high-density scenes, our co-occurrence prior based on density information cannot find more faces, but our method does not deteriorate the original performance in low-density scenarios.

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. Co-occurrence prior can detect more true faces and makes sense to detect low-resolution faces in the crowded challenge. S^3NMS avoids arbitrary discarding or preservation of the bounding box and reduces false positives and increases true positives. The proposed method does not require any extra training and is simple to implement. In the future, we will further explore richer context information to solve low-resolution face detection in crowded 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. Tychsen-Smith and L. Petersson. Improving object localization with fitness nms and bounded iou loss. In *CVPR*, pages 6877–6885, 2018.

[30] L. Tychsensmith and L. Petersson. Improving object localization with fitness nms and bounded iou loss. In *CVPR*, pages 6877–6885, 2018.

[31] L. Wolf and S. M. Bileschi. A critical view of context. In *IJCV*, pages 251–261, 2006.

[32] T. Wu, D. Liang, J. Pan, and S. Kaneko. Context-anchors for hybrid resolution face detection. In *ICIP*, pages 3297–3301, 2019.

[33] 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.

[34] 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.

[35] S. Yang, P. Luo, C. C. Loy, and X. Tang. Wider face: A face detection benchmark. In *CVPR*, pages 5525–5533, 2016.

[36] Y. Yoo, D. Han, and S. Yun. Extd: Extremely tiny face detector via iterative filter reuse. In *arXiv*, 2019.

[37] S. Zhang, X. Zhu, and Z. Lei. S^3fd: Single shot scale-invariant face detector. In *ICCV*, pages 192–201, 2017.

[38] 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.

[39] 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.

[40] 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.

[41] C. Zhu, R. Tao, K. Luu, and M. Savvides. Seeing small faces from robust anchor's perspective. In *CVPR*, pages 5127–5136, 2018.

[42] 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.

[43] 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.

[44] Y. Zhu, H. Cai, S. Zhang, C. Wang, and Y. Xiong. Tinaface: Strong but simple baseline for face detection. In *arXiv*, 2020.