Multimedia Tools and Applications (2022) 81:35411–35430
https://doi.org/10.1007/s11042-022-12319-y

1204: MULTIMEDIA TECHNOLOGY FOR SECURITY AND SURVEILLANCE IN DEGRADED VISION

Inferred box harmonization and aggregation for degraded face detection in crowds

Dong Liang[1] · Qixiang Geng[1] · Han Sun[1] · Huiyu Zhou[2] · Shun'ichi Kaneko[3]

Received: 29 December 2020 / Revised: 26 August 2021 / Accepted: 17 January 2022 /
Published online: 1 April2022

Abstract
Since objects usually keep a certain distance from the surveillance camera, small object detection is a practical issue. Detecting small objects is also one of the remaining challenges in the computer vision community. The current detectors usually leverage a more robust backbone network, build one or more multi-scale feature pyramids, or define a more precise anchor-box screening criteria. However, the distinguishable features are scarce due to the appearance degradation and a shallow resolution. In this paper, we leverage high-level context to enhance anchor-based detectors' capabilities for small and crowded face detection. We first define face co-occurrence prior based on density maps (FCP-DM) to explore extensive high-level contextual information. We propose a score-size-specific non-maximum suppression (S^3NMS) to replace the traditional non-maximum suppression at the end of anchor-based detectors. Our approach is plug and play and model-independent, which could be concatenated into the existing anchor-based face detectors without extra learning. Compared to the prior art on the WIDER FACE hard set, our method increases an Average Precision of 0.1%-1.3%, while on Crowd Face, which we make for testing small and crowded face detection, it raises an Average Precision of 1% - 6%. Codes and dataset have been available online.

Keywords Object detection · Degraded face · Video surveillance

1 Introduction

For video surveillance in the open world, robust face detection is an ultimate component to handle various facial-related tasks. Since the faces are usually far from the surveillance

✉ Dong Liang
liangdong@nuaa.edu.cn

1 College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China

2 School of Informatics, University of Leicester, Leicester LE1 7RH, UK

3 Metatec Cooperation, Yokohama 220-0004, Japan

camera, small face detection is a problem with practical needs. In recent years, renewed detection paradigms [8, 13], strong backbone [14, 15, 20] and large-scale datasets [9, 31] jointly push forward the limit of face detection to approach humans' cognition. However, because flexible mechanisms and abundant domain knowledge guide human's cognition, human has advantages on handing the challenges of low-resolution [18]. In the computer vision community, a central issue of small object detection is the appearance degradation of a small object with shallow resolution. The essential issue is that the distinguishable features are scarce due to appearance degradation.

Anchor-based face detectors have achieved satisfactory performance on the benchmark WIDER FACE [31]. Recently, many face detectors rely on features extracted from deep Convolutional Neural Network (CNN). They obtain low-level features of the objects (such as texture and edge related feature) from the low layers of the network and high-level features such as semantic-related feature from the high layers. However, for face detectors, thorny issues involved in detecting degraded faces are caused by small-size, defocus blur, and occlusion [35]. These blurred and low-resolution faces only have dozens or even a few pixels, containing limited feature information. When using the standard spatial pooling process [35] in a CNN, appearance features would be further degraded. This problem is ill-posed for a low-resolution object as CNN can only provide very few low-level features at the low layers and almost no high-level features of these faces at the high layers. Therefore, aggregating more information from context becomes a inevitable choice.

Some works [8, 20, 30, 35, 38] have introduced contextual information for low-resolution face detection. In these methods, the contextual information of faces is usually employed in the form of low-level context via an augmented receptive field of feature maps. Obviously, rich low-level context is helpful to detect small objects and easy to implement [3, 23], but augmented receptive field relies on the limited local area. Some current detectors leverage a more robust backbone network [13], build one or more multi-scale feature pyramids [8], or define a more precise anchor-box screening criteria [20]. On the other hand, [1] shows that humans detecting objects that violate their high-level context take longer and make more errors. Hence, object detection is expected to fit into a certain high-level scene context to reach humans' cognition.

We also argue that high-level contextual information is valuable for small object detection. Different from the traditional context which rely on adjusting the local receptive fields, we explore the compositional semantics–the relationship among the confidence, quantity and size of objects, as high-level contextual information and extend it to the whole scene. We presents a universal strategy with density-map-based face co-occurrence priors (FCP-DM) and score-size-specific non-maximum suppression (S^3NMS), independent of training paradigms to directly replace the standard non-maximum suppression (NMS) post-processing formula in anchor-based detectors. FCP-DM harmonizes the outputs of a detector according to crowd density estimation. It enhances the sensitivity and specificity of the detector via increasing true positives. S^3NMS aggregates the bounding box by decreasing false positives and increasing true positives according to the inferred face boxes' score and size. Figure 1 illustrates the proposed detection framework. We also collect a challenging face detection dataset with tiny faces to provide adequate samples to further prominent the bottleneck of detecting crowded faces.

The contribution of this paper are listed as follows.

- We proposed a general approach using high-level contextual information for small and crowd face detection.

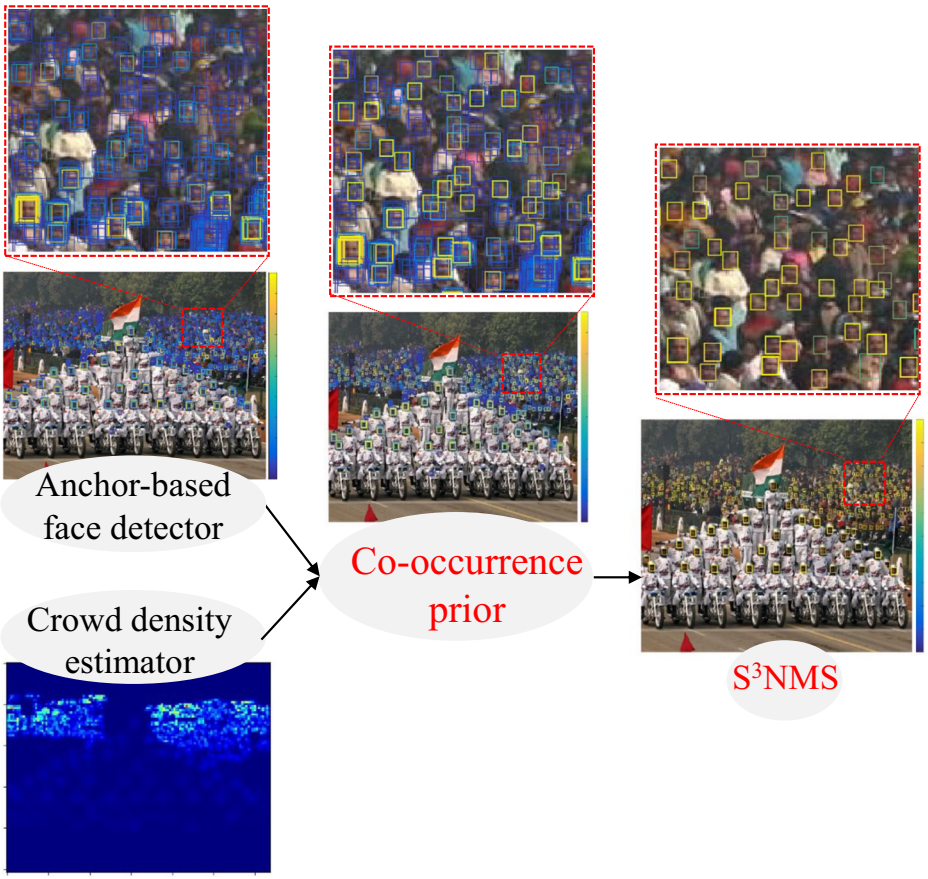


Fig. 1 Architecture of the proposed framework. Face co-occurrence priors increase true positives of the inferred face boxes according to crowd density estimation. S^3NMS further increases true positive and reduces false positive according to the inferred face boxes' score and size. Detector confidence is given by the color bar on the right of the images, blue boxes represent low confidence, and yellow boxes represent high confidence

- The proposed scheme reduces false positives and increases true positives according to the inferred face boxes' score, quantity, and size under the guidance of crowd density estimation.
- The proposed scheme makes sense to detect multi-scale and low-resolution faces in the crowded challenge and provides a refined structure to avoid arbitrary discarding or preservation of the bounding box.
- It requires no extra training and is simple to be implemented.

The remainder of this paper is organized as follows. We discuss the related work in Section 2. We describe problem formulation and the proposed FCP-DM and S^3NMS in detail in Section 3. The experimental results are presented and discussed in Section 4, and the conclusions, limitations, and future work are presented in Section 5.

2 Related work

2.1 Anchor-based object detection model

Supervised training of a detection model requires bounding boxes and their class labels associated with the objects in images. However, it is not trivial for CNNs to directly predict an order-less set of arbitrary cardinals [24, 32]. One commonly used strategy is to introduce anchors, which employ a divide-and-conquer strategy to match objects with convolutional features spatially. Anchor box is firstly introduced in Faster R-CNN [20] and serves as a reference at multiple scales and aspect ratios for object detection. During inference, anchors independently predict object bounding boxes, where the box with the highest classification score is retained after the non-maximum suppression (NMS) procedure. Anchor-based detection methods include the well-known FPN [13], RetinaNet [14], SSD [15], and YOLOv3 [19], all of which requires additional post-processing, i.e. NMS. Anchor-free approaches, including CornerNet [10], CenterNet [4], and ExtremeNet [37] et al., have shown a great potential for the cases of extreme object scales and aspect ratios. However, without the anchor box as the reference point, direct regression of bounding boxes from convolutional features remains challenging. On the benchmark WIDER FACE [31], the most competitive methods are still the anchor-based models. We continue to tap the potential of anchor-based methods, expecting to enhance these methods' performance without additional training.

2.2 Using context in face detection

The idea of using context in object detection has been studied in many works. Divvala et al., Oliva and Torralba and Wolf and Bileschi [3, 18, 28] reviewed contextual information used in contemporary methods and analysed its role for challenging object detection in empirical evaluation. For specific face detection, Hybrid Resolution Model (HR) [8] is a simple yet effective framework for finding small faces, demonstrating that both large context and scale-variant representations are crucial. 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, [38] pools ROI features around faces and bodies for detection, which improves overall performance. The methods mentioned above either build multi-scale feature pyramids or enlarge feature maps' receptive fields to employ the low-level context. But for the task of small face detection, the above methods are inefficient in using context information and not as flexible as a human cognitive system. For the human cognitive system, high-level context and domain knowledge help reduce decision time and disambiguate the low-quality inputs. We expect to fit into a proper high-level context of a scene to enhance the anchor-based face detectors.

2.3 Inferred box harmonization

The goal of Non-Maximum Suppression (NMS) [21] penalizes false positive detections, which has been an integral part of many object detection algorithms in computer vision [7, 17, 22, 27]. Soft-NMS in [2] argues that the conventional NMS is too greedy because only the bounding box with the maximum score is selected. In contrast, all other bounding boxes with a significant overlap with this box are suppressed using a pre-defined threshold. Soft-NMS suppresses the bounding box by reducing its score instead of just removing it. In our preliminary experiments, however, soft-NMS causes the increase of false positives because

some redundant boxes cannot be deleted due to high scores. More complex learning-based methods rely on the model-related learning process. Hosang [6] proposed a learning-based NMS to improve localization and occlusion handling. Tychsen-Smith [26] 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. Although learning-based methods have achieved good performance in specific scenarios, they also have poor generalization capabilities and insufficient cross-domain adaptability. We tend to develop plug-and-play and model-independent paradigms, which could be integrated into the existing anchor-based detectors without extra learning.

3 Problem formulation and proposed approaches

3.1 Co-occurrence prior based on density maps

3.1.1 Crowd density map

A density map is firstly used in crowd counting literature. Zhang et al. [36] proposes geometry adaptive and fixed kernels with Gaussian convolution to generate a density map. Li et al. [12] introduces a dilated convolutional neural network to improve the density map's quality. Liu et al. [16] 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 our problem formulation, crowd density estimation is employed to derive face co-occurrence priors for harmonizing a face detector's outputs.

A density map is also used in crowd analysis since it can exhibit the headcount, locations and their spatial distribution. Given a set of N training images $\{I_i\}_{(1\le i\le N)}$ with corresponding ground-truth density maps D_i^{gt}, the goal of density map estimation is to learn a non-linear mapping $\mathscr{F}$ that maps an input image I_i to an estimated density map $D_i^{est}(I_i) = \mathscr{F}(I_i)$, 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 $P_i^{est} = \left\{P_{i,j}\right\}_{1\le j\le C_i}$ 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 $\mathscr{N}^{est}\left(p \mid \mu, \sigma^2\right)$. We have

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

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 $\{d_k^{i,j}\}_{(1\le k\le K)}$. 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, since the individuals are close to each other in a high-density crowd scene, it 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) [16]. It combines features obtained using multiple receptive field sizes and learns the importance of each feature at each image location. It adaptively encodes the scale of the contextual information required to predict crowd density accurately. This method yields an algorithm that outperforms state-of-the-art crowd counting methods, especially when perspective effects are strong.

3.1.2 Co-occurrence of homogeneous faces for inferred box harmonization

In this part, we focus on using the face co-occurrence prior to optimize the detectors in crowd scenarios. Since the face size approaches the limit of imaging resolution, the face appearance is scarce and inadequate. A low-resolution face that is difficult for humans to recognize is also a challenge for a vision-based detector. General face detectors are highly dependent on appearance features, and the severe scarcity of information is essentially ill-posed, which can directly lead to the degradation of detection performance. However, it is unavoidable normality in crowd-scene face detection. We utilize the co-occurrence of faces as a higher-level context to make more sensitive detection when the face is ambiguously or marginally visible in a crowd scene.

Face co-occurrence prior here refers to the harmonization of homogeneous faces–If the scores of many faces dominate in an image, it is reasonable that some inferred boxes 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.

We send the image into the density estimate network to generate the density map D_i^{est} first. From the perspective of making full use of the context, the contextual information on a broader perception area of a density map could provide more co-occurrence prior to the area just around the observed face. Hence, it is unreasonable to use (2) to estimate face co-occurrence directly. However, using density maps to reconcile the results of face detection seems to be a chicken-egg paradox. At least, how to use the inaccurate density map to adjust the result of face detection is a complicated interaction problem of heterogeneous information.

As mentioned earlier, the head size in a high-density crowd scene can be represented by a density map rather than in a low-density crowd scene. Hence, we need to design an operator to disturb the inference in high-density areas and give up interventions for low-density areas. We define a dense grid on image I_i, and generate blocks $A = \{A_i^n\}$ with 50% overlapping to minimize border effects, where n is the number of blocks. The population in different blocks is estimated by integrating over the values of the predicted density map as follows,

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

In the corresponding block, the average size of all the high score faces is calculated and recorded as BS_{avg}^n.

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

where a_i^n is the area of region A_i^n. There are two constraints to filter the inferred box for reconciliation. If the score of a inferred box $s_{x,y}$ exceeds the score threshold s_t, the inferred box could be a candidate of human face. The inferred boxes whose scores are ultimately lower than s_t will be deleted. These boxes with the size between $(1-\gamma, 1+\gamma)BS_{avg}$ are further filtered out as the inferred box for reconciliation. The reconciliation formula is as

follows,

$$s_{x,y} = \sigma[D^{est}_{i(x,y)}(p \mid A^n_i)]s_{x,y} + s_{x,y}, \tag{5}$$

where σ is the Sigmoid function. The above proposed FCP-DM scheme is summarized in Algorithm 1.

Algorithm 1 Face co-occurrence prior for inferred box Harmonization.

Data: $\mathcal{B} = \{b_{x,y}\}, \mathcal{S} = \{s_{x,y}\}, A = \{A^n_i\}, D^{est}_{i(x,y)}, \gamma, s_t,$
B is the list of initial inferred boxes, S contains corresponding inferred scores, A^n_i is the list of different density areas, D^{est}_i 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); z_i^n ← Σ_{p∈A_i^n} D_est(p|A_i^n);
        if s_{x,y} ⩾ s_t then
            m ← m + 1; BS_sum ← BS_sum + BS_i
        BS_avg ← BS_sum/m
for b_{x,y} in B do
    if b_j in A_j then
        if (1 − γ)BS_avg ≤ BS_{x,y} ≤ (1 + γ)BS_avg and s_{x,y} ⩾ s_t then
            s_{x,y} = σ[D^{est}_{i(x,y)}(p | A_i^n)]s_{x,y} + s_{x,y}
```

3.2 Score-Size-specific NMS

NMS [21] is utilized as standard 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, the Intersection over Union overlap (IoU) refers to the standard Jaccard similarity 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. This principle is also effective in the multi-scale pyramid scheme, as more inferred boxes may be detected in different pyramid layers. However, this will cause missed detection, as the face covered by part of another face or two faces close to each other may not be detected. As illustrated in Fig. 2, in the process of the three models moving close to each other, the middleman's face was gradually covered, and the detection score decreased significantly. Meanwhile, due to the instability fluctuation of the maximum value (as shown in Fig. 2), the use of NMS will aggravate the instability of the detection score.

Based on NMS, soft-NMS [2] provides a chance to preserve the overlapped and closed objects using a penalizing function to the inferred scores. NMS is a non-continuous procedure to produce a penalty when an IoU threshold of N_t is reached, which could lead to abrupt changes to the ranked score list of the inferred boxes. A continuous penalty function should have no penalty when there is no overlap and a large penalty at a high overlap. Also, when the overlap is low, it should gradually increase the penalty, and b_u should not affect

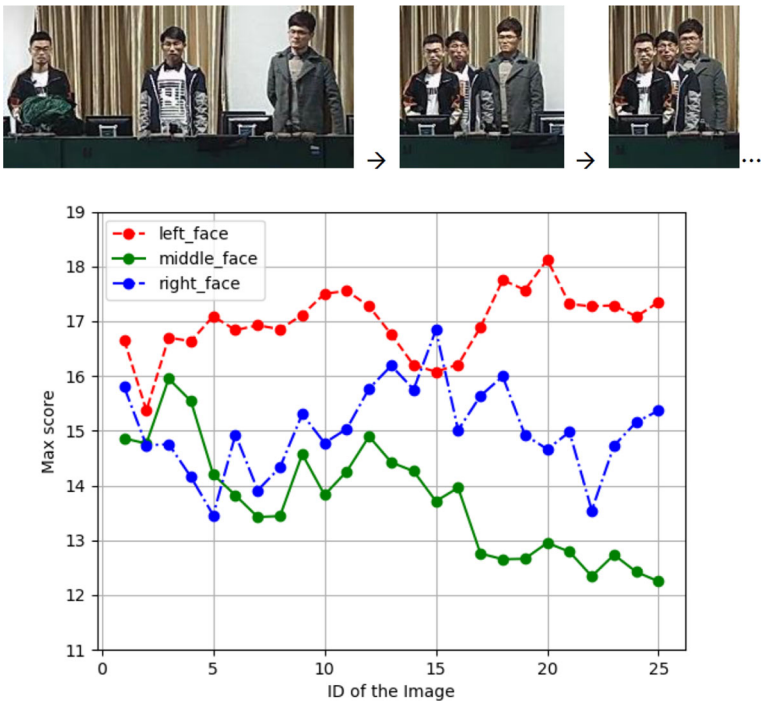


Fig. 2 The statistic of the maxim detection score of HR [8] face detection model using a Hikvision surveillance camera. In the process of the three models moving close to each other, the middleman's face was gradually covered, and the detection score decreased significantly.

the scores of boxes with very low overlap. However, when the overlap of a box b_v with b_u becomes close to 1, b_v should be significantly penalized. Taking this into consideration, 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 the scores of all the remaining detection boxes are updated. It suppresses the inferred box by reducing its score instead of just removing it. However, Both NMS and soft-NMS ignore the role of size factor in the inferred box aggregation. Consider an extreme situation, the areas of the two boxes are quite different, that is $b_u >> b_v$. From the definition of (6), the intersection is much smaller than the union. 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, a more reasonable way should be to implement a retention operator among similar size boxes. Based on IoU, we define ACB (Area Consistency of boxes) as follows,

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

We adopt a constraint that $ACB(b_u, b_v)$ must be in a range between $1 \pm B_t$. Algorithm 2 summarizes the proposed S^{3}NMS scheme. If $IoU(b_m, b_{x,y}) \geqslant N_t$ and $1 - B_t \leqslant ACB(b_m, b_{x,y}) \leqslant 1 + B_t$, where b_m is the box with the highest score in $\mathscr{B}$, 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 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. The above scheme gives a chance to detect faces covered by other faces without causing false positives as soft-NMS does.

Algorithm 2 Score-size-specific NMS.

Data: $\mathcal{B} = \{b_{x,y}\}, \mathcal{S} = \{s_{x,y}\}, N_t, S_{th}, B_t$
$\mathcal{B}$ is the list of initial inferred boxes, $\mathcal{S}$ contains corresponding inferred scores, N_t is the IoU threshold, S_{th} is the score threshold, B_t is the IoB threshold.
for $b_{x,y}$ *in* $\mathcal{B}$ **do**
  $b_m \leftarrow argmax(\mathcal{S})$
  if $IoU(b_m, b_{x,y}) \geqslant N_t$ *and* $1 - B_t \leqslant ACB(b_m, b_{x,y}) \leqslant 1 + B_t$ **then**
    if $s_{x,y} \geqslant S_{th}$ **then**
      $s_{x,y} = s_{x,y}e^{-\mathrm{IoU}(b_u,b_v)^2/\delta}$
    else
      $\mathcal{B} \leftarrow \mathcal{B} - b_{x,y}; \mathcal{S} \leftarrow \mathcal{S} - s_{x,y}$

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.

4 Experimental evaluation

4.1 Dataset preparation

In face detection literature, a widely used benchmark is WIDER FACE [31]. 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, including size changes, occlusion, pose changes, lighting changes, and background confusion.

Considering the proposed solution in this paper is mainly for obscured small face detection in crowd scenes, in addition to the commonly used WIDER FACE, we prepare a new dataset - Crowd Face by ourselves collected from the Internet. There are 34 images with 10731 annotated faces, and the maximum number of faces on an image is 1001. As illustrated in Fig. 3, we measured the average size of objects (blue plots) and the average number of objects per image (orange plots). Crowd Face has much smaller (around 10 times smaller in the average size of objects) and more faces (approximately 20 times more in average number of objects per image) than WIDER FACE. As shown in Fig. 5 and Appendix B, the Crowd Face dataset has many low-resolution, small, and obscured faces. It is a chal-

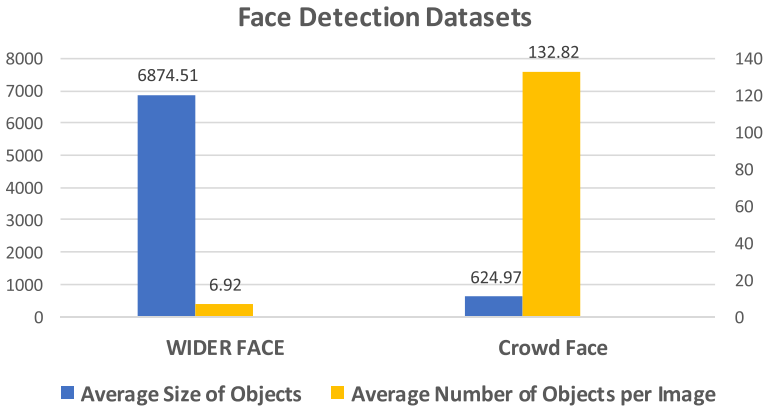


Fig. 3 Comparison of benchmark dataset WIDER FACE and our Crowd Face dataset. Two quantities are measured for each dataset: average size of objects (blue plots) and average number of objects per image (orange plots)

lenging dataset with hard samples, specifically for high-density face detection. Testing face detection algorithms on Crowd Face is helpful to explore the shortages of face detectors.a

4.2 Experimental setting

In our experiments, the models we used to verify our proposed methods are HR [8], EXTD [33], S^{3}FD [34], LFFD [5], CAHR [29], PyramidBox [25], DSFD [11] and TinaFace [39]. 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 FCP-DM, $S_{th} = 0.5$, $B_t = 0.1$ for S^{3}NMS. Our experiments are run on GTX1080 with 16 GB RAM and 12-core i7 CPU.

4.3 Experiments for face co-occurrence prior based on density map

In this part, the co-occurrence priors based on a density map is tested on Crowd Face, as the proposed method is mainly used to detect faces in high-density crowd scenes. We introduce density information to the state-of-the-art anchor-based detectors, and then combine with our proposed algorithm. As is illustrated in Fig. 4, we integrate FCP-DM to the trained detectors: HR [8], CAHR [29], EXTD [33], S^{3}FD [34], PyramidBox [25] and DSFD [11], and compare their performance with the original detectors. The red curve in each figure represents our proposed co-occurrence prior based on the density map integrated into the detector. The blue curve below represents the original detector without our proposed method. The results show that the proposed FCP-DM has higher accuracy at the same precision rate than the original detectors. Face co-occurrence priors increase true positives according to crowd density estimation. Figure 5 shows the comparison of the co-occurrence prior within HR (cyan ellipses) and original HR (magenta rectangles) in crowd scenes, where the proposed approach detects more true faces. It illustrates that the proposed method can enhance the detectors to find more true faces in crowd scenes with many low-resolution small faces.

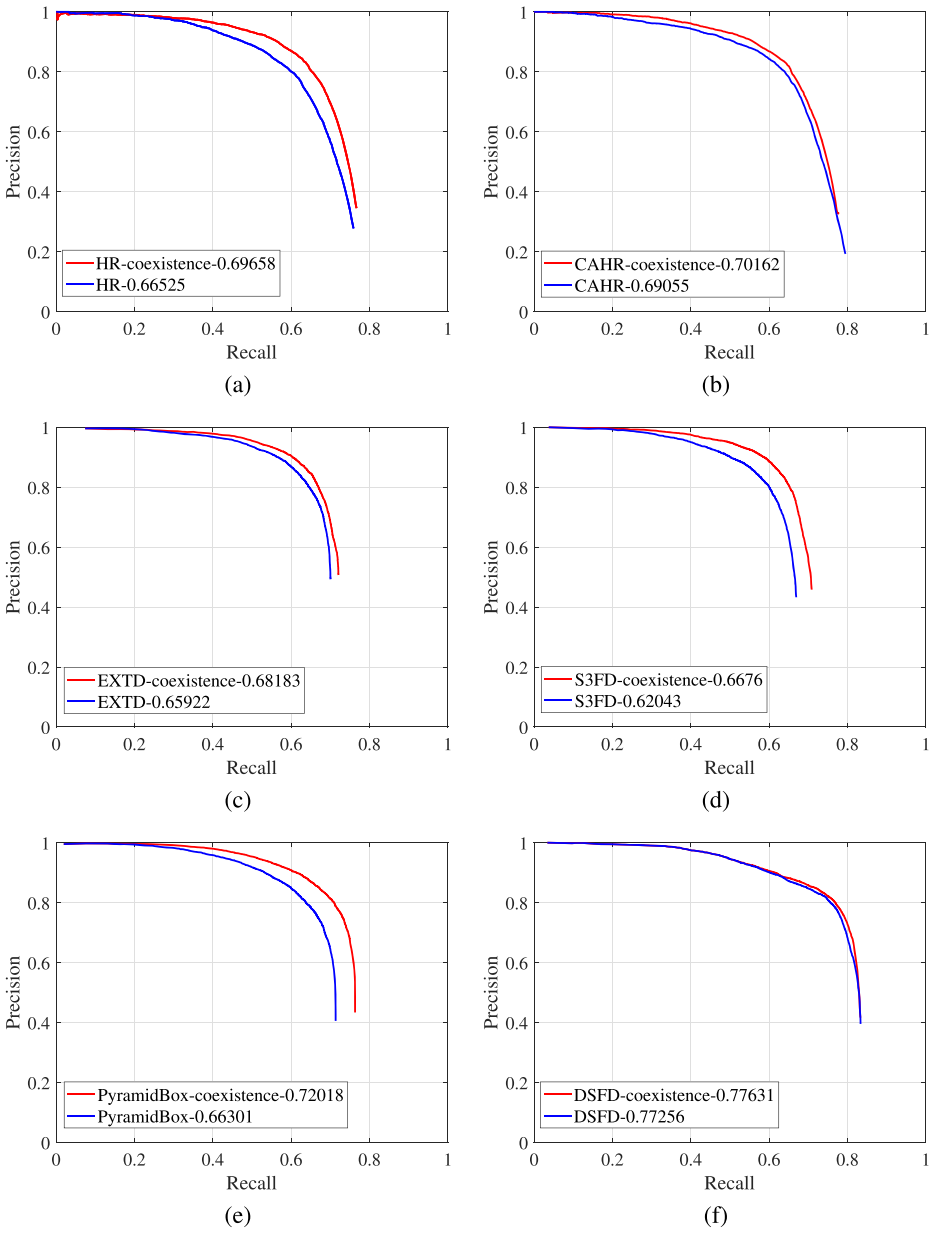


Fig. 4 P-R curve of FCP-DM, compared with the original models (HR [8], CAHR [29], EXTD [33], S^3FD [34], PyramidBox [25], DSFD [11])

4.4 Experiments for score-size-specific NMS

In this part, the score-size-specific NMS (S^3NMS) is tested. The test sets include the WIDER FACE hard set and the Crowd Face, the test models include LFFD [5], HR [8],

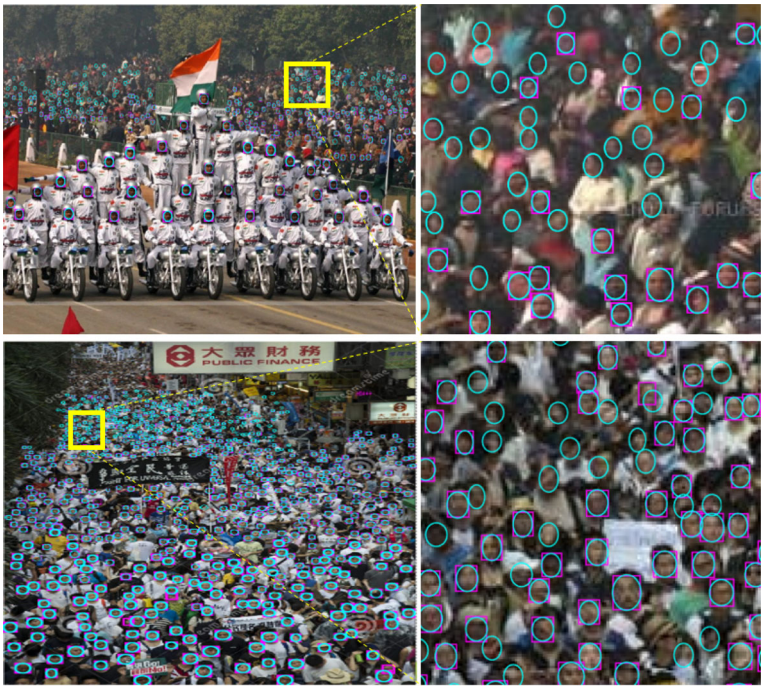


Fig. 5 A comparison of low-resolution face detection in Crowd Face Dataset using our proposed method within HR detector [8] (cyan ellipses) and the original HR (magenta rectangles)

S3FD [34], CAHR [29], PyramidBox [25], EXTD [33], DSFD [11], the NMS threshold is 0.3.

We integrate NMS, soft-NMS, and our proposed S^3NMS into the above state-of-the-art detectors; these detectors are all anchor-based models. Our proposed S^3NMS is a post-processing method without any additional training. We compared our approach with other post-processing methods NMS and Soft-NMS, which do not need model training too, as shown in Table 1, S^3NMS has the highest 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. Figure 6 shows the comparison of true and false positives for original HR, CAHR, PyramidBox, and these models with our

Table 1 AP performance of NMS, Soft-NMS and 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
	0.665	0.683	0.707	HR
Crowd Face	0.691	0.707	0.720	CAHR
	0.663	0.671	0.681	PyramidBox

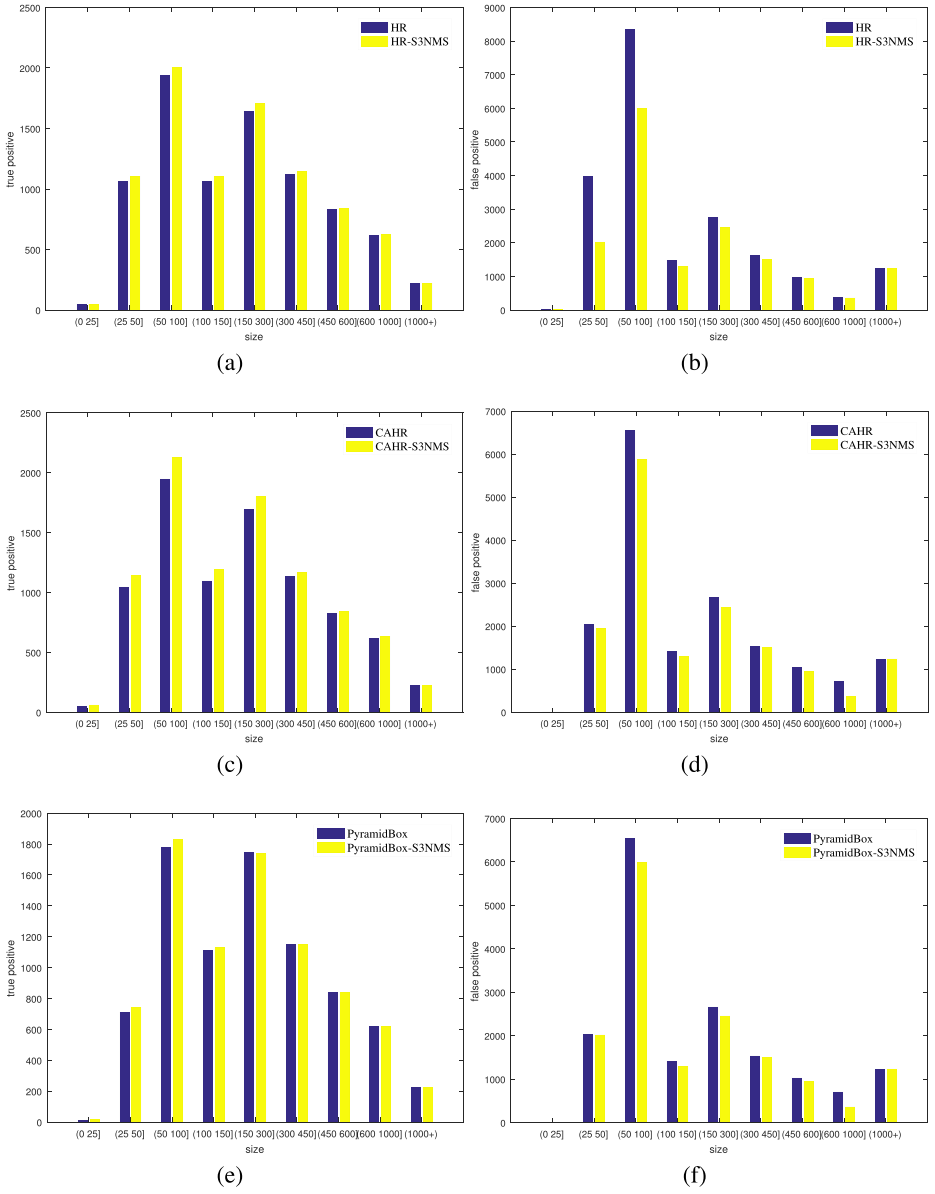
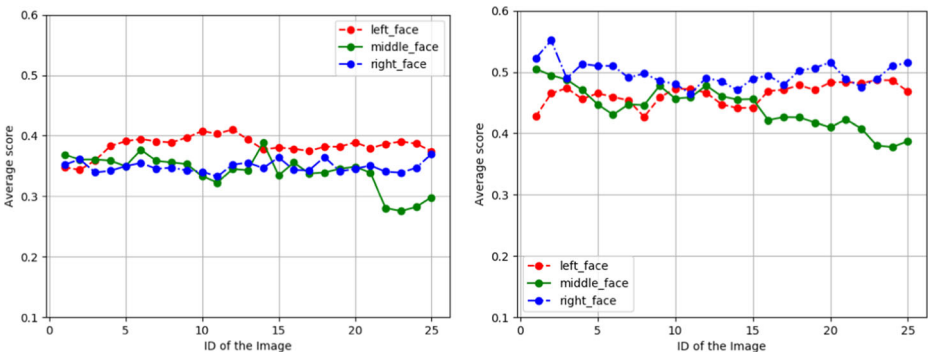


Fig. 6 Comparison of true and false positives for original HR [8], CAHR [29], PyramidBox [25], and these models with the proposed score-size-specific NMS

proposed S^3NMS on Crowd Face. The right yellow plots in each figure represent our proposed S^3NMS integrated with the detector, and the left blue plots represent NMS integrated with the detector. Figure 7 shows the statistic of the average detection score of HR [8] before and after cascading with the proposed score-size-specific NMS. The average scores for the 3 models (shown in Figure 2) have been significantly improved after cascading with S^3NMS.



(a) HR detector (b) HR cascading with S^3NMS

Fig. 7 The statistic of the average detection score of HR [8] before and after cascading the proposed score-size-specific NMS. The experimental scene is the same as Fig. 2

4.5 Ablation study on crowd face

As shown in Table 2, we perform ablation experiments on Crowd Face. We separately integrate NMS, score-size-specific NMS, and co-occurrence prior based on density maps to HR [8], PyramidBox [25], EXTD [33], CAHR [29], and DSFD [11] on Crowd Face. We first compare the performance of NMS and our proposed S^3NMS, which shows that our proposed S^3NMS has higher AP performance. Then, we respectively integrate NMS and S^3NMS with FCP-DM into the detectors. The result shows the proposed FCP-DM can further improve the performance, and S^3NMS combined with FCP-DM has the best AP performance. It shows that our proposed S^3NMS has higher AP performance than NMS, and S^3NMS combined with co-occurrence priors has higher AP performance than integrates only one of the two methods into the detectors. FCP-DM and S^3NMS increase an overall AP around 1% - 6% for the above detection models.

4.6 Overall performance on WIDER FACE

In this part, we apply the proposed FCP-DM and S^3NMS to the detectors together on the WIDER FACE dataset. The trained detectors are LFFD [5], HR [8], CAHR [29], EXTD [33], S^3FD [34], PyramidBox [25], DSFD [11], and TinaFace [39] and compare their performance with the original detectors. Table 3 shows that the proposed approach integrating within most face detectors has better performance than the original methods. It illustrates that our proposed score-size-specific NMS reduces false positives and increases true positives according to the inferred face boxes' score and size. For WIDER FACE hard set, our results could increase an AP of 0.1-1.3%, indicating the capability of the proposed approach in challenging situations. The WIDER FACE-easy set contains almost no high-density scenes—the proposed FCP-DM freezes the density estimation output. Thus the proposed method maintains consistent results in most models and does not deteriorate the original performance in low-density scenarios. More demonstration with degraded images are shown in Appendix A, B, and C.

Table 2 Ablation study of our proposed score-size-specific NMS and co-occurrence priors based on density maps to HR, PyramidBox, EXTD, CAHR, DSFD and TinaFace on Crowd Face

Method	NMS	S^3NMS	Coexist.	$AP(\%)$
	✓			0.665
HR		✓		0.707
[8]	✓		✓	0.697
		✓	✓	0.710
	✓			0.663
PyramidBox		✓		0.681
[25]	✓		✓	0.720
		✓	✓	0.725
	✓			0.659
EXTD		✓		0.674
[33]	✓		✓	0.682
		✓	✓	0.688
	✓			0.691
CAHR		✓		0.720
[29]	✓		✓	0.702
		✓	✓	0.728
	✓			0.772
DSFD		✓		0.780
[11]	✓		✓	0.776
		✓	✓	0.781
	✓			0.771
TinaFace		✓		0.776
[39]	✓		✓	0.781
		✓	✓	0.784

Table 3 Performance of integrating score-size-specific NMS and co-occurrence priors 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 [29]	0.928	0.928	0.912	0.913	0.832	0.844
EXTD [33]	0.921	0.923	0.911	0.912	0.846	0.853
S^3FD [34]	0.945	0.945	0.934	0.936	0.853	0.855
PyramidBox [25]	0.960	0.960	0.948	0.950	0.888	0.890
DSFD [11]	0.966	0.966	0.957	0.957	0.905	0.906
TinaFace [39]	0.963	0.964	0.956	0.958	0.930	0.932

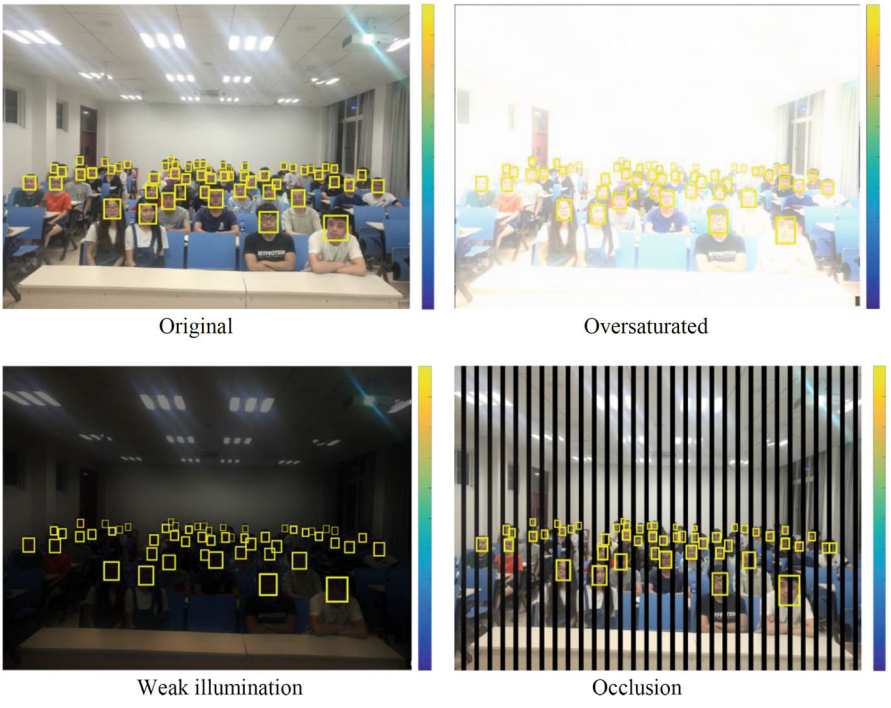
5 Conclusion, limitations, and future work

We proposed a general approach using high-level contextual information for small and crowd face detection. The proposed scheme reduces false positives and increases true positives according to the inferred face boxes' score, quantity, and size under the guidance of crowd density estimation. The proposed scheme makes sense to detect multi-scale and low-resolution faces in the crowded challenge and provides a refined structure to avoid arbitrary discarding or preservation of the bounding box. It requires no extra training and is simple to be implemented.

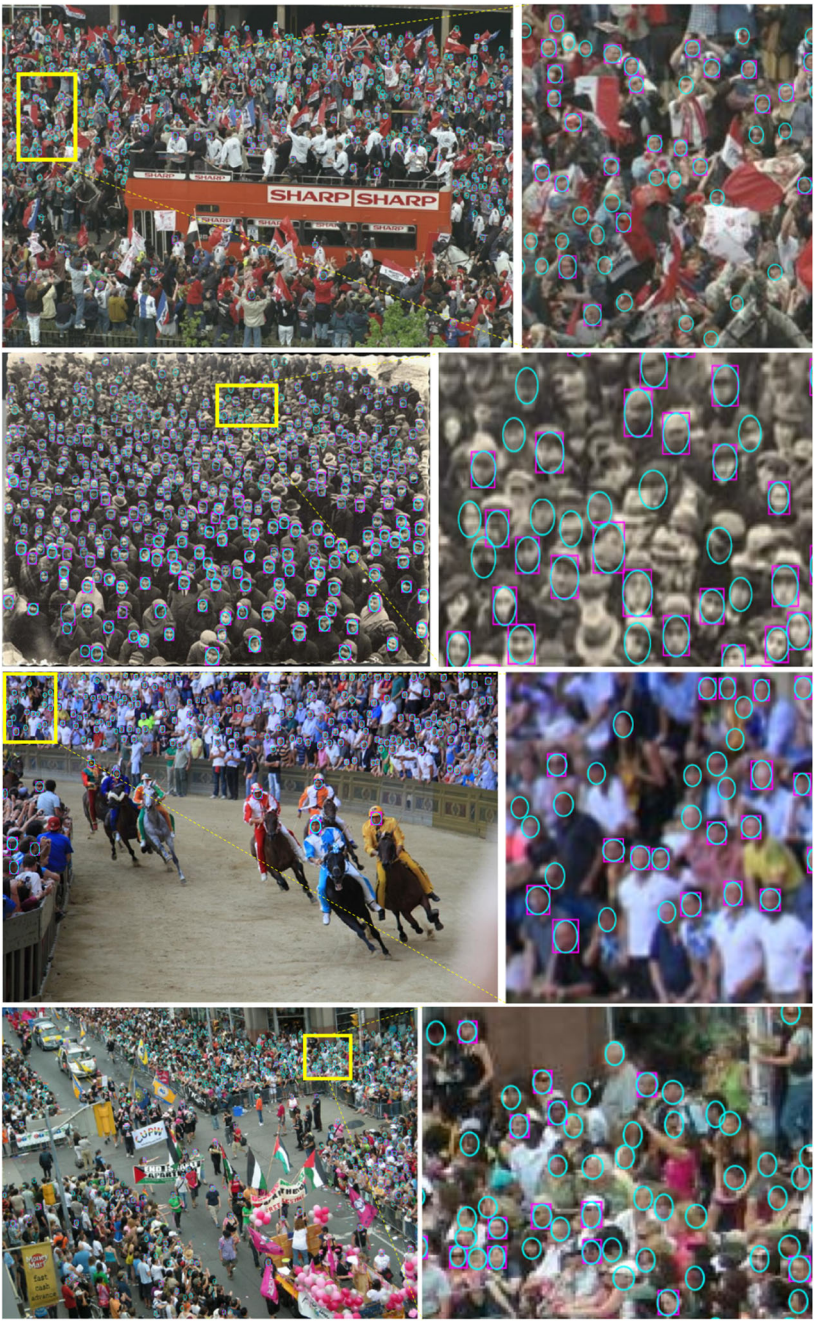
The main limitation of this method is that it needs to rely on the performance of the density estimate network. The performance of this network is usually affected by the quality of training data, scene category, and task category. Therefore, the network will need to be retrained when changing specific tasks, such as vehicle density estimation.

We will explore the capability of the proposed framework for other dense and small object detection tasks, such as remote sensing scenes with rotated bounding boxes.

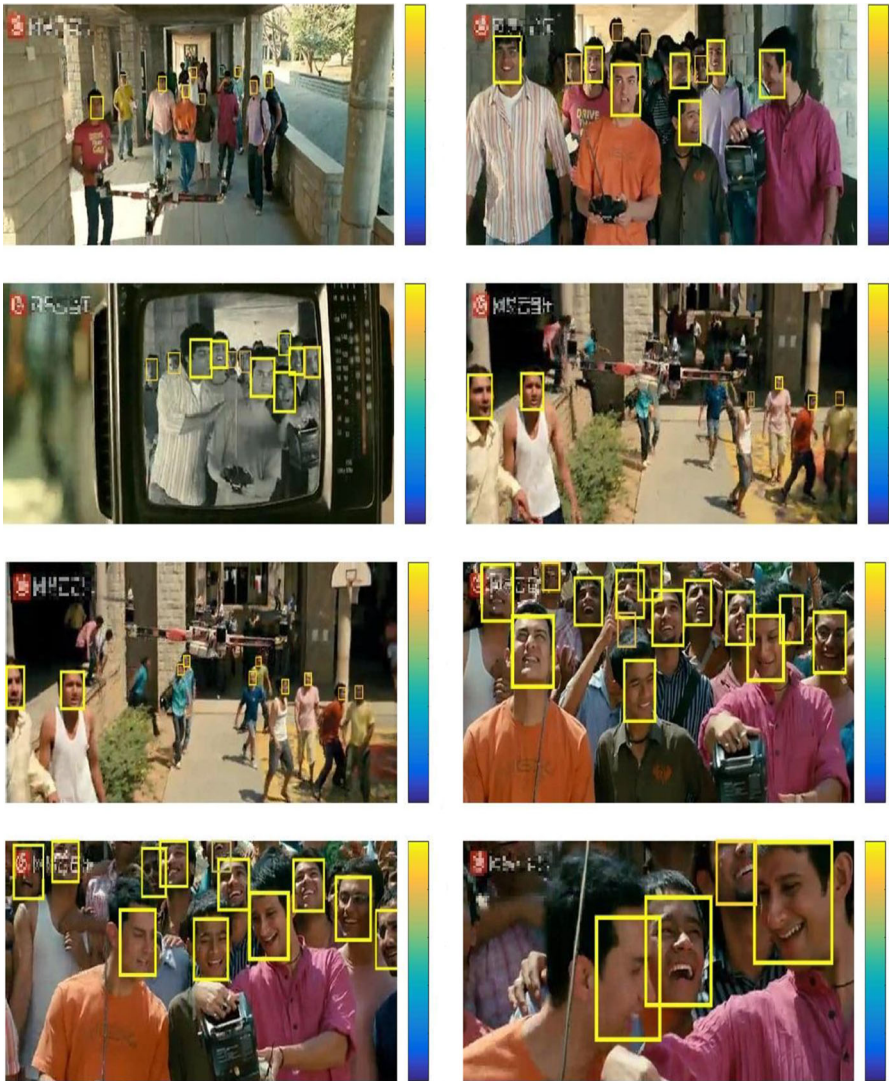
Appendix A: Results in a series of degraded images in a classroom in NUAA (All the portrait rights are licensed)



Appendix B: Comparing the proposed scheme in the HR detector (cyan ellipses) and the standard NMS scheme in the HR detector (magenta rectangles) on Crowd Face



Appendix C: Performance in various open scenes using "Three Idiots" movie clip



References

1. Biederman I, Mezzanotte RJ, Rabinowitz JC (1982) Scene perception" detecting and judging objects undergoing relational violations. Cogn Psychol 14(2):143–177
2. Bodla N, Singh B, Chellappa R, Davis LS (2017) Soft-nms improving object detection with one line of code. International Conference on Computer Vision
3. Divvala SK, Hoiem DW, Hays J, Efros AA, Hebert M (2009) An empirical study of context in object detection. IEEE Conf Comput Vis Pattern Recognit pp 1271–1278

4. Duan K, Bai S, Xie L, Qi H, Huang Q, Tian Q (2019) Centernet: Keypoint triplets for object detection. International Conference on Computer Vision pp 6569–6578
5. He Y, Xu D, Wu L (2019)
6. Hosang J, Benenson R, Schiele B (2017) Learning non-maximum suppression. IEEE Conf Comput Vis Pattern Recognit pp 4507–4515
7. Hosang JH, Benenson R, Schiele B (2016) A convnet for non-maximum suppression. German Conference on Pattern Recognition pp 192–204
8. Hu P, Ramanan D (2017) Finding tiny faces. IEEE Conference on Computer Vision and Pattern Recognition 1522–1530
9. Jain V, Learned-Miller E (2010) Fddb: A benchmark for face detection in unconstrained settings. Tech. Rep. UM-CS-2010-009, University of Massachusetts Amherst
10. Law H, Deng J (2018) Cornernet: Detecting objects as paired keypoints. European Conference on Computer Vision pp 734–750
11. Li J, Wang Y, Wang C (2019) Dsfd: Dual shot face detector. IEEE Conference on Computer Vision and Pattern Recognition pp 5060–5069
12. Li Y, Zhang X, Chen D (2018) Csr-net: Dilated convolutional neural networks for understanding the highly congested scenes. IEEE Conference on Computer Vision and Pattern Recognition pp 1091–1100
13. Lin T, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: 2017 IEEE Conference on computer vision and pattern recognition (CVPR), pp 936–944
14. Lin T, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 42(2):318–327
15. Liu W, Anguelov D, Erhan D, Szegedy C, Reed SE, Fu C, Berg AC (2016) SSD: Single Shot multibox detector. European Conference on Computer Vision pp 21–37
16. Liu W, Salzmann M, Fua P (2019) Context-aware crowd counting. IEEE Conf Comput Vis Pattern Recognit pp 5099–5108
17. Neubeck A, Van Gool L (2006) Efficient non-maximum suppression. International Conference on Pattern Recognition 3:850–855
18. Oliva A, Torralba A (2007) The role of context in object recognition. Trends Cogn Sci 11(12):520–527
19. Redmon J, Farhadi A (2018)
20. Ren S, He K, Girshick RB, Sun J (2017) Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149
21. Rosenfeld A, Thurston M (1971) Edge and curve detection for visual scene analysis. IEEE Trans Comput 20(5):562–569
22. Rothe R, Guillaumin M, Van Gool L (2014) Non-maximum suppression for object detection by passing messages between windows. Asian Conference on Computer Vision pp 290–306
23. Shrivastava A, Gupta A, Girshick RB (2016) Training region-based object detectors with online hard example mining. IEEE Conf Comput Vis Pattern Recognit pp 761–769
24. Stewart R, Andriluka M, Ng AY (2016) End-to-end people detection in crowded scenes. In: CVPR, pp 2325–2333
25. Tang X, Du DK, He Z (2018) Pyramidbox: a context-assisted single shot face detector. European Conference on Computer Vision pp 797–813
26. Tychsen-Smith L, Petersson L (2018) Improving object localization with fitness nms and bounded iou loss. IEEE Conf Comput Vis Pattern Recognit pp 6877–6885
27. Tychsensmith L, Petersson L (2018) Improving object localization with fitness nms and bounded iou loss. IEEE Conf Comput Vis Pattern Recognit pp 6877–6885
28. Wolf L, Bileschi SM (2006) A critical view of context. Int J Comput Vis 69(2):251–261
29. Wu T, Liang D, Pan J, Kaneko S (2019) Context-anchors for hybrid resolution face detection. In: 2019 IEEE International conference on image processing, pp 3297–3301
30. Xiang W, Zhang D, Yu H, Athitsos V (2018) Context-aware single-shot detector. Workshop on Applications of Computer Vision pp 1784–1793
31. Yang S, Luo P, Loy CC, Tang X (2016) Wider face: a face detection benchmark. IEEE Conf Comput Vis Pattern Recognit pp 5525–5533
32. Yang T, Zhang X, Li Z, Zhang W, Sun J (2018) etaanchor: Learning to detect objects with customized anchors. In: In neurIPS, pp 320–330
33. Yoo Y, Han D, Yun S (2019)
34. Zhang S, Zhu X, Lei Z (2017) S^3fd: Single shot scale-invariant face detector. International Conference on Computer Vision pp 192–201
35. Zhang S, Zhu X, Lei Z, Wang X, Shi H, Li SZ (2018) Detecting face with densely connected face proposal network. Chinese Conference on Biometric Recognition 284:3–12

36. Zhang Y, Zhou D, Chen S, Gao S, Ma Y (2016) Single-image crowd counting via multi-column convolutional neural network. IEEE Conf Comput Vis Pattern Recognit pp 589–597
37. Zhou X, Zhuo J, Krahenbuhl P (2019) Bottom-up object detection by grouping extreme and center points. IEEE Conf Comput Vis Pattern Recognit pp 850–859
38. Zhu C, Zheng Y, Luu K, Savvides M (2017) Cms-rcnn: Contextual multi-scale region-based cnn for unconstrained face detection. Deep learning for biometrics pp 57–79
39. Zhu Y, Cai H, Zhang S, Wang C, Xiong Y (2020) Tinaface: Strong ut simple baseline for face detection. In: Arxiv