Accepted Manuscript

Detecting Small Faces in the Wild Based on Generative Adversarial Network and Contextual Information

Yongqiang Zhang, Mingli Ding, Yancheng Bai, Bernard Ghanem

PII: S0031-3203(19)30198-0
DOI: <https://doi.org/10.1016/j.patcog.2019.05.023>
Reference: PR 6917

To appear in: *Pattern Recognition*

Received date: 23 January 2019
Revised date: 28 March 2019
Accepted date: 13 May 2019

Please cite this article as: Yongqiang Zhang, Mingli Ding, Yancheng Bai, Bernard Ghanem, Detecting Small Faces in the Wild Based on Generative Adversarial Network and Contextual Information, *Pattern Recognition* (2019), doi: <https://doi.org/10.1016/j.patcog.2019.05.023>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Highlights

- A novel unified end-to-end convolutional neural network architecture for small face detection is proposed.
- A regression branch is introduced to the GAN-based architecture for further refining the locations of small faces in the wild.
- New losses are designed to train the GAN-based network for small face detection in the wild.
- Contextual information around face regions is further utilized to detect hard faces in the real-world scenarios.
- The performance of our method outperforms previous state-of-the-art approaches by a large margin on WIDER FACE dataset, especially on the most challenging Hard subset.

Detecting Small Faces in the Wild

Based on Generative Adversarial Network and Contextual Information

Yongqiang Zhang^a, Mingli Ding^{a,*}, Yancheng Bai^{b,*}, Bernard Ghanem^b

^a*School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin 15001, China*

^b*Visual Computing Center, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia*

Abstract

Face detection techniques have been developed for decades, and one of the remaining open challenges is detecting small faces in unconstrained conditions. The reason is that tiny faces are often lacking detailed information and blurry. In this paper, we proposed an algorithm to directly generate a clear high-resolution face from a small blurry one by adopting a generative adversarial network (GAN). Toward this end, the basic GAN formulation achieves it by super-resolving and refining sequentially (*e.g.* SR-GAN and Cycle-GAN). However, we design a novel network to address the problem of super-resolving and refining jointly. Moreover, we also introduce new training losses (*i.e.* classification loss and regression loss) to promote the generator network to recover fine details of the small faces and to guide the discriminator network to distinguish face *vs.* non-face and to refine location simultaneously. Additionally, considering the importance of contextual information when detecting tiny faces in crowded cases, the context around face regions is combined to train the proposed GAN-based network for mining those very small faces from unconstrained scenarios. Extensive experiments on the challenging datasets WIDER FACE and Fddb demonstrate the effectiveness of the proposed method in restoring a clear high-resolution face from a small blurry one, and show that the achieved performance outperforms previous state-of-the-art methods by a large margin.

Keywords: Face detection, Tiny faces, Super-resolution, Generative adversarial network, Contextual information

*Corresponding author.

Email address: mingli.ding.hit@gmail.com (Mingli Ding)

1. Introduction

Face detection is a fundamental and important problem in computer vision, since it is usually a key step towards many subsequent face-related applications, including face parsing [1, 2], face verification [3, 4], face tagging and retrieval [5, 6], *etc.* . Meanwhile, face detection has been widely studied over the past few decades and numerous accurate and efficient methods have been proposed for most constrained scenarios. Recently, some works try to detect faces in uncontrolled settings, which is much more challenging due to the significant variations in scale, blur, pose, expressions and illumination. A thorough survey on face detection methods can be found in [7].

Modern face detectors have achieved impressive results on large and medium scale faces. However, the performance on small faces is far from satisfactory. The main difficulty for small face (*e.g.* 10×10 pixels) detection is that small faces usually lack sufficient detailed information to distinguish them from similar backgrounds (*e.g.* regions of partial faces or hands). Another problem is that modern CNN-based face detectors use down-sampled convolutional (*conv*) feature maps with stride 8, 16 or 32 to represent faces, which lose most spatial information and are too coarse to describe small faces. To detect those small faces, [8] directly up-samples images using bi-linear operation and exhaustively searches faces on the up-sampled images. However, this method will increase the computation cost and the inference time will increase significantly too. Moreover, images are often zoomed in with a small upscaling factor ($2\times$ at most) in [8], otherwise, artifacts will be generated. Besides, [9, 10, 11] use intermediate *conv* feature maps to represent faces at specific scales, which keeps the balance between the computation burden and the performance. However, the shallow but fine-grained intermediate *conv* feature maps lack discrimination information, which causes many false positive results. More importantly, these methods take no care of other challenges, like blur and illumination, *etc.* .

To deal with the nuisances in face detection, we propose a unified end-to-end convolutional neural network for better face detection based on the classical generative adversarial network (GAN) framework. There are two sub-networks in our detector, a generator network and a discriminator network. In the generator network, a super-resolution network (SRN) is used to up-sample small faces to a fine scale for finding those tiny faces. Compared to re-sizing by bi-linear operation, SRN can reduce artifacts and improve the quality of up-sampled images with a large upscaling factor ($4\times$ in our current implementation), as shown in Figure 1 (b) and

Figure 1: The detection results of tiny faces in the wild. (a) is the original low-resolution blurry face, (b) is the result of re-sizing directly by a bi-linear kernel, (c) is the generated image by the super-resolution method, our result (d) is learned by the super-resolution ($\times 4$ upscaling) and refinement network simultaneously, and (e) is our result when adding the regression loss and the contextual information. Note that the green bounding-boxes are the detection results from our method in [12], and red bounding-boxes are the new findings from the detection network driven by the added regression loss and contextual information. Best viewed in color and zoomed in.

(c). However, even with such sophisticated SRN, up-sampled images are unsatisfactory (usually blurry and lacking fine details) due to faces of very low resolutions (*e.g.* 10×10 pixels). Therefore, a refinement sub-network (RN) is proposed to recover some missing details in the up-sampled images and generate sharp high-resolution images for classification, as shown in Figure 1 (d). In the discriminator network, we introduce a new loss function that enforces the discriminator network to distinguish the real/fake face and face/non-face simultaneously. Finally, the generated images and real images pass through the discriminator network to JOINTLY distinguish whether they are real high-resolution images or generated super-resolved images and whether they are faces or non-faces. More importantly, the classification loss is further used to guide the generator network to generate clearer faces for easier classification.

As we all know, detecting small blurry faces in unconstrained conditions is very difficult. Even though the approximate location of a face is found, it sometimes fails to meet the strict evaluation metrics (*i.e.* Intersection over Union (IoU) with the annotated bounding-box > 0.5),

especially for some super small and blurry faces. For small faces detection, a few pixel drift may lead the predicted bounding-boxes fail to fulfill the above mentioned evaluation criteria. To overcome this issue, we further add a regression branch in the discriminator network to refine the predicted bounding-boxes, as shown in Figure 2. Meanwhile, the regression loss is also back-propagated to the generator network, which encourages the generator to produce higher quality images for easier classification and better localization.

Due to the complexity of detecting tiny faces in the wild, the proposed face detector based on the above mentioned GAN may miss some hard faces (*i.e.* the faces are too far from the camera, *e.g.* 10×10 pixels, and too blurry) or may introduce some false positive results (*i.e.* some very similar backgrounds or parts of the face, like the nose, are predicted as a face). To address these problems, we further combine the face region with the contextual information around faces to train our proposed GAN-based network for finding those very small faces and alleviating those false positive results. Actually, the context plays a significant role in some computer vision tasks [13], like tracking, pedestrian detection, object detection, *etc.* . As we all know, faces never appear isolated in the real-world, and they always occur simultaneously with the neck, shoulders, or body, which provide lots of contextual information to be exploited especially when the faces are not distinguishable for low-resolution, or blurry cases. In our implementation, we adopt the method of adding proportional contexts, *i.e.* enlarging the crop window by X2.0 when preparing the training data, as shown in Figure 4. Those face regions cropped by the enlarged window own rich contextual information, and experiments prove that it is helpful for finding tiny faces and distinguishing hard faces from similar backgrounds. Figure 1 (e) shows a face generated from the generator network when adding the regression loss and the contextual information. Moreover, the red bounding-boxes in Figure 1 are some hard cases, which are missed when the regression loss and the contextual information are not introduced in the networks.

Contributions. To sum up, this paper makes the following four main contributions. **(1)** A novel unified end-to-end convolutional neural network architecture for face detection is proposed, where super-resolution and refinement network are used to generate real and sharp high-resolution images and a discriminator network is introduced to classify faces *vs.* non-faces. **(2)** New losses are introduced to guide the discriminator network to distinguish real/fake images, faces/non-faces, and to refine face locations simultaneously. More importantly, the classification and regression

losses are back-propagated to promote the generator network to generate clearer faces with fine details for easier classification and better localization. **(3)** Contextual information around face regions is further utilized to train the proposed GAN-based network, which is helpful for detecting tiny faces and distinguishing hard faces from similar backgrounds. **(4)** Finally, we demonstrate the effectiveness of our proposed method in restoring a clear high-resolution face from a small blurry face, and show that the detection performance outperforms the previous state-of-the-art approaches by a large margin on WIDER FACE dataset, especially on the most challenging Hard subset.

This paper is an extended version of our CVPR18 oral paper [12]. In particular, we make three methodological improvements: the first one is that we change the backbone network of the discriminator from VGG16 [14] to ResNet50 [15], which can learn stronger features for those small blurry faces, and can be trained in an end-to-end way with the generative network. To justify the power of the replaced backbone network, we conduct an experiment on these two networks as shown in Table 2, and we observe that the AP performance improves from 94.4%/93.3%/87.3 to 94.6%/93.6%/87.6% on the Easy/Medium/Hard subset respectively. The second one is that we add a regression branch in the discriminative network for further refining the predicted bounding-boxes of each target face. Meanwhile, the regression loss is further back-propagated to the loss function of the generator network to generate high quality super-resolved images for easier classification and better localization. The ablation studies in subsection 4.3 demonstrate the effectiveness of the regression branch and the regression loss. The third one is that contextual information around face regions is combined to train the proposed GAN-based face detection network, and the ablation study experiments validate the positive influence of introducing the contextual information to find those very small blurry faces. In addition, we provide more results to demonstrate the effectiveness of our proposed method, and we also conduct more experiments to support our viewpoint. For instance, we present how sensitive our proposed method is to the baseline face detector by conducting additional experiments.

The rest of the paper is organized as follows. We review the related work in Section 2. In Section 3, the detailed architecture of our GAN-based face detector, the objective function of the proposed network, and the method of combining contextual information are described. In Section 4, we present some experiments and ablation studies on two widely used face detection benchmarks (*i.e.* WIDER FACE and FDDB). Finally, the conclusions and further work are

provided in Section 5.

2. Related Work

2.1. Face Detection

As a classic topic, numerous face detection systems have been proposed during the past decades or so. Existing face detection methods can be broadly categorized as handcrafted feature based methods [16, 17, 18, 19] and CNN-based methods [20, 21, 10, 11, 22, 23]. However, most of detection systems based on handcrafted features only train a single scale model, which is applied to each level of a feature pyramid, thus increasing the computation cost drastically, especially for the complicated features. Moreover, the limited representation of handcrafted features restricts the performance of detectors, particularly in uncontrolled settings.

Inspired by the great success of Faster RCNN, several recent works [10, 11] utilized this framework to detect faces and showed impressive performance on Fddb benchmark [24]. However, the performance drops dramatically on more challenging WIDER FACE dataset [25], which contains a large number of faces with lower resolution. The main reason for this disparity is that deep *conv* feature maps with lower spatial resolution are used as final features, which is insufficient for handling small faces [20, 21]. To overcome this problem, detectors [10, 11] have to upsample by re-sizing input images to different scales during training and testing, which inevitably increases the memory and computation costs. Furthermore, the generated images by these re-sized methods often have large structural distortions, as shown in Figure 1 (b). Compared to these methods, our method exploits the super-resolution and refinement networks to generate clear and fine faces with high resolution ($4\times$ up-scaling), as shown in Figure 1 (d) and (e), and then a discriminator network is trained to distinguish faces from the input images.

2.2. Super-resolution and Refinement Network

With the development of deep learning, great improvements have been achieved on super-resolution [26, 27, 28]. However, when obtaining these promising results, there is a precondition that the down-sampling kernel is known, and most of these CNN-based super-resolution methods can not be applied to uncontrolled settings (*i.e.* in the wild).

There are different refinement networks for different tasks, and the most similar refinement method to our refinement network is the methods of deblurring. Most existing deblur methods

heavily rely on prior models to solve this ill-posed problem, and a prior assumes that gradients of natural images have a heavy-tailed distribution [29, 30]. Recently, CNNs have also been used to deblur the low-quality images [31, 32, 33, 34]. However, these deblurring methods still involve explicit kernel estimation, and the recovered images usually have significant ringing artifacts if the estimated kernels are inaccurate.

Although existing super-resolution methods and refinement methods are effective at up-sampling and refining images respectively, it is not easy to extend to jointly super-resolving and refining the low-resolution images. [8] proposed a method to simultaneously reconstruct a clear high-resolution image from a blurry low-resolution input. However, their blurry low-resolution images are obtained by using the bicubic interpolation down-sampling and a known blur kernel from the high-resolution images (*i.e.* their blurry low-resolution images are synthetic). In this paper, we design a novel network to generate a clear super-resolution face from a small blurry one which is collected from the real-world. We would like to note that our work is the first trying to jointly super-resolve and refine the small blurry faces in the wild.

2.3. Generative Adversarial Network

In the seminal work [35], generative adversarial network (GAN) is introduced to generate realistic-looking images from random noises. GANs have achieved impressive results in image generation [36], image editing [37], representation learning [38], image super-resolving [39] and character transferring [40]. Recently, GAN has been applied to super-resolution (SRGAN) [39] and has obtained promising results. Compared to super-resolution on natural images, face images in the wild are of arbitrary poses, illumination and blurring, so super-resolution on face images is much more challenging. More importantly, the high resolution images generated by SRGAN are blurry and lack fine details especially for low-resolution faces, which are unfriendly for the face classifier. Towards this end, we design a refinement sub-network to recover some detailed information. In the discriminator network, the basic GAN [39, 40, 35] are trained to distinguish real and fake high resolution images. To classify faces or non-faces, we extend the discriminator network to classify fake *vs.* real images, faces *vs.* non-faces, and regress locations of target faces simultaneously. Furthermore, the classification and regression losses are further propagated back to the generator network, and guide the generator network to reconstruct the clearer super-resolution images for easier classification and better localization.

2.4. Context-associated Methods

Contextual information has been paid much attention to in some computer vision tasks, like tracking [41], pedestrian detection [42], object detection [43, 44], *etc.*, and it is usually a critical factor to mine small instances in those tasks. Here, we only talk about the context-assisted methods for small object detection and tiny face detection. In object detection, [43] uses spatial recurrent neural networks to integrate contextual information outside the region of interest for detecting the small objects. [45] encodes the contextual information by a multi-region deep convolutional neural network and shows a large improvement on small object detection. [46] proposes an object relation module to model the relations of a set of objects through interaction between their appearance feature and geometry. [44] proposes a context refinement algorithm and explores rich contextual information to better refine each proposed region in two-stage object detectors. In face detection, contextual information has also been widely used, especially for finding small and blurry faces in the unconstrained conditions. CMS-RCNN [11] imitates human vision system and allows explicit body contextual reasoning in the network for detecting the small faces. [47] explores three problems in the context of detecting tiny faces, (*i.e.* the role of scale in-variance, image resolution, and contextual reasoning), and shows context is crucial for finding tiny faces in the wild. SSH [48] includes a simple context module which models the contextual information by increasing the effective receptive field on the convolutional layers. [13] proposes an anchor-based context-assisted method to introduce supervised information on learning contextual features for detecting small, blurred and partially occluded faces.

Motivated by the success of using contextual information to find the small faces in literature [47, 13], we also combine the context around face regions with the face region itself to detect the tiny faces in the unconstrained settings. Different from [47, 13], which encode the contextual information on the feature maps and rely on some unique mechanisms. However, we do not need these difficult procedures and we only explore the contextual information in the stage of preparing the training data (*i.e.* by enlarging the crop window), which is simpler and more efficient when comparing with the above mentioned methods.

3. Proposed Method

In this section, we will introduce our proposed method in details. First, we revisit the baseline MB-FCN detector, and give a brief description on the classical GAN network. Then, the

Figure 2: The architecture of the proposed tiny face detector system, which includes generator network and discriminator network. (A) The generator network is trained to reconstruct a clear super-resolution image ($4\times$ upscaling) from the low-resolution input image, which includes an upsample sub-network and a refinement sub-network. (B) The discriminator network is a ResNet50 architecture with three parallel fc layers, and the first fc layer is to distinguish the natural real high-resolution images or the generated super-resolution images, the second one is to classify faces or non-faces, and the third fc layer is to refine the location of target faces.

whole architecture of our proposed method is presented, as shown in Figure 2. In addition, we introduce each part of our proposed networks in details and define the loss functions for training the generator network and discriminator network respectively. Finally, we show the method of utilizing contextual information to find hard faces and alleviate false positive results.

3.1. Revisit MB-FCN Detector

We adopt a multi-branch fully-convolutional network (MB-FCN) face detector [9] as our baseline detector to generate the training/testing data (*i.e.* faces and non-faces) for training/testing our proposed generator network and discriminator network, as shown in Figure 3. During testing, the candidate faces (*i.e.* region of interest, RoI) are cropped by the MB-FCN detector, and those ROIs pass through the generator network to produce the clear super-resolution faces. At the same time, the discriminator network produces the detection results (*i.e.* classify faces *vs.* non-faces, and refine predicted bounding-boxes).

Figure 3: Illustration of the baseline MB-FCN detector, and the pipeline of the method for generating the training/testing data. (A) The input images are fed into the network; (B) MB-FCN [9] detector is our baseline, it crops the positive data (*i.e.* faces) and negative data (*i.e.* non-faces) from the input images for training the proposed generator network and the discriminator network, or generates the regions of interests (ROIs) for testing. Note that both cropped face and non-face patches are inputted into the generator network for super-resolution and refinement. (C) The positive data and negative data (or ROIs) are generated by the MB-FCN detector. Best viewed on the computer, in color and zoomed in.

The whole architecture of MB-FCN detector consists of four components. (i) The first component is the shared intermediate *conv* layers, which can be of any typical architecture like AlexNet [49], VGGNet [14] or ResNet [15]. We employ ResNet50 [15] pre-trained on the ImageNet dataset as the backbone network to learn convolutional feature maps. (ii) The second component creates the skip connections of feature maps at different scales to represent targets of different resolutions, where up-sampling and down-sampling operates on the convolutional feature maps of different scales to produce K_{th} branch feature maps. (iii) For each branch, multi-task learning is deployed to learn one fully-connected network (FCN) to deal with faces within certain scales. Specifically, each FCN takes as input a 1×1 spatial window of the input convolutional feature maps, and outputs a lower-dimensional feature, which is fed into two sibling fully-connected layers (*reg* and *cls*) with filter size 1×1 . (iv) The *reg* and *cls* outputs of each branch are converted to scored bounding boxes. Then, non-maximum suppression (NMS) is applied to those bounding-boxes with confidence above a predefined threshold, and the final face detection results are obtained. More details can be found in [9].

3.2. GAN

GAN [35] learns a generative model G via an adversarial process, and it trains a generator network G and a discriminator network D simultaneously. The training process alternately optimizes the generator and discriminator, which compete with each other. The generator G is trained for generating the samples to fool the discriminator D , and the discriminator D is trained to distinguish real images and fake images generated from the generator. The objective function can be defined as follows:

$$\mathcal{L}_{GAN}(G, D) = \mathbb{E}_{x \sim p_{data}(x)}[\log D_{\theta}(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D_{\theta}(G_{\omega}(z)))], \quad (1)$$

where z is the random noise and x denotes the real data, θ and ω denote the parameters of G and D respectively. Here, G tries to minimize the objective function and adversarial D tries to maximize it as Eq(2):

$$\arg \min_G \max_D \mathcal{L}_{GAN}(G, D). \quad (2)$$

Similar to [35, 39], we further design a generator network G_{w_G} which is optimized in an alternative method along with a discriminator network D_{θ_D} to solve the small face super-resolution, object classification, and bounding box regression problems simultaneously. The overall loss is defined as follows:

$$\arg \min_{w_G} \max_{\theta_D} \mathbb{E}_{(I^{HR}, y, v) \sim p_{train}(I^{HR}, y, v)}[\log D_{\theta_D}(I^{HR}, y, v)] + \mathbb{E}_{(I^{LR}, y, v) \sim p_G(I^{LR}, y, v)}[\log(1 - D_{\theta_D}(G_{w_G}(I^{LR}, y, v)))], \quad (3)$$

where I^{LR} denotes face candidate regions with low-resolution, I^{HR} represents the face candidate regions with high-resolution, y is the label (*i.e.* face or non-face), and v is the ground-truth bounding-box regression target. Unlike [35], the input of our generator is low-resolution images rather than the random noise. Different from [39], we up-sample and refine the input image simultaneously in the generator network. In the discriminator network, we have multi-task, where we distinguish the generated *vs.* true high-resolution images, classify faces *vs.* non-faces, and regress locations of target faces jointly.

3.3. Network Architecture

Our generator network includes two components (*i.e.* upsample sub-network and refinement sub-network), and the first sub-network takes low-resolution images (*i.e.* including faces and

Table 1: Architecture of the generator and discriminator network. “conv” represents a convolutional layer, “x8” denotes a residual block which has 8 convolutional layers, “de-conv” means a fractionally-stride convolutional layer, “2x” denotes up-sampling by a factor of 2, “L.” denotes the layer, and “fc” indicates a fully connected layer.

	Generator												Discriminator					
	Up-sample Sub-network						Refinement Sub-network											
Layer	conv	conv x8	conv	de-conv	de-conv	conv	conv	conv x8	conv	conv	conv	conv	conv	L.1	L.2	L.3	L.4	fc
Kernel Num.	64	64	64	256	256	3	64	64	64	256	256	3	64	128	256	512	1024	2
Kernel Size	3	3	3	3	3	1	3	3	3	3	3	3	3	1	1	1	1	-
Stride	1	1	1	2	3	1	1	1	1	1	1	1	2	1	2	2	2	-

non-faces patches) as inputs and the outputs are super-resolution images. Since small blurry faces lack fine details and due to the influence of MSE loss Eq(4), the generated super-resolution faces are usually blurring. So we design the second sub-network to refine the super-resolved images generated from the first sub-network. Furthermore, we add a classification branch to the discriminator network for the purpose of face detection, which means our discriminator can classify faces and non-faces as well as distinguish fake and real images. Meanwhile, we further introduce a regression branch in the discriminative network to refine the predicted bounding-boxes for detecting those hard faces.

Generator network (G_w). As shown in Table 1 and Figure 2, we adopt a deep CNN architecture, which has shown effectiveness for image super-resolution in [39], to build our generator network. There are two fractionally-strided convolutional layers [50] (*i.e.* de-convolutional layer) in the network, and each de-convolutional layer consists of learned kernels which perform up-sampling a low-resolution image to a $2\times$ super-resolution image. In contrast to their networks, our generator network includes a refinement sub-network which is also a deep CNN architecture. Similar to [50], we use batch-normalization [51] and rectified linear unit (ReLU) activation after each convolutional layer except the last layer.

The up-sampling sub-network first up-samples a low-resolution image and outputs a $4\times$ super-resolution image, which is usually blurring when the inputted small faces are far from the cameras or under fast motion. Then, the refinement sub-network processes this blurry image, and outputs a clear super-resolution image, which is easier for the discriminator to classify faces *vs.* non-faces, and to regress locations of target faces.

Discriminator network (D_θ). We employ ResNet50 [15] as our backbone network in the discriminator network, as shown in Table 1. To avoid too many down-sampling operations for small blurry faces, we remove the max-pooling layer from the “conv5” layer. Moreover, we replace all fully connected layers (*i.e.* $fc6$, $fc7$, $fc8$) with three parallel fully connected layers fc_{GAN} , fc_{cls} and fc_{reg} . For the specific tasks, the adversarial fc layer fc_{GAN} and classification layer fc_{cls} use the sigmoid loss function [50], while the regression layer fc_{reg} uses the smooth $L1$ loss function as in [52].

The input of discriminator network is super-resolution RoI images which include natural real high-resolution images and generated fake super-resolution images, the output of fc_{GAN} branch is the probability of the input being a real image, the output of fc_{cls} branch is the probability of the input being a face, and the output of fc_{reg} branch is the bounding-box regression offsets ($t = (t_x, t_y, t_w, t_h)$) for the ROI candidate images.

3.4. Loss Function

We adopt the pixel-wise loss and adversarial loss as some previous approaches [39, 40] to optimize our generator network. In contrast to [39], we remove the VGG feature matching loss due to the calculation cost and we introduce the classification and regression losses to drive the generator network to recover fine details from blurry small faces.

Pixel-wise loss. The input of our generator network is small blurry images instead of the random noise [35]. A natural way to enforce the outputs of the generator to be close to the super-resolution ground truths is through the pixel-wise MSE loss, and it is calculated as Eq(4):

$$L_{MSE}(w) = \frac{1}{N} \sum_{i=1}^N (||G_{1_{w_1}}(I_i^{LR}) - I_i^{HR}||^2 + ||G_{2_{w_2}}(G_{1_{w_1}}(I_i^{LR})) - I_i^{HR}||^2), \quad (4)$$

where I^{LR} and I^{HR} denote small blurry images and super-resolution images respectively, G_1 means up-sampling sub-network, G_2 denotes the refinement sub-network and w is the parameters of generator network. However, while achieving less loss between the generated and neutral high-resolution images in pixel level, the solution of the MSE optimization problem usually lacks the high-frequency content which results in perceptual unsatisfactory images with over-smooth texture. Also, this is one of the reasons why the generated super-resolution images are blurry.

Adversarial loss. To achieve more realistic results, we introduce the adversarial loss [39] to the objective loss, defined as Eq(5):

$$L_{adv} = \frac{1}{N} \sum_{i=1}^N \log(1 - D_{\theta}(G_w(I_i^{LR}))), \quad (5)$$

where θ denotes the parameters of the discriminator network D . Here, the adversarial loss encourages the network to generate sharper high-frequency details for trying to fool the discriminator network. In Eq(5), the $D_{\theta}(G_w(I_i^{LR}))$ is the probability of the reconstruction image $G_w(I_i^{LR})$ being a natural super-resolution image.

Classification loss. In order to make the reconstructed images by the generator network easier to classify, we also introduce the classification loss to the objective loss. Let $\{I_i^{LR}, i = 1, 2, \dots, N\}$ and $\{I_i^{HR}, i = 1, 2, \dots, N\}$ denote small blurry images and high-resolution real natural images respectively, and $\{y_i, i = 1, 2, \dots, N\}$ represents the corresponding labels, where $y_n = 1$ or $y_n = 0$ indicates the image is the face or non-face respectively. The formulation of classification loss is like Eq(6):

$$L_{cls} = \frac{1}{N} \sum_{i=1}^N (\log(y_i - D_{cls}(G_w(I_i^{LR}))) + \log(y_i - D_{cls}(I_i^{HR}))). \quad (6)$$

Our classification loss plays two roles, where the first one is to distinguish whether the high-resolution images, including both generated and natural real high-resolution images, are faces or non-faces in the discriminator network. The other role is to promote the generator network to reconstruct a sharper image with fine details.

Regression Loss. To enable more accurate localization for some hard faces, we also introduce a bounding-box regression loss to the objective function, formulated as Eq (7):

$$L_{reg}(t, v) = \frac{1}{N} \sum_{i=1}^N \sum_{j \in \{x, y, w, h\}} [u_i \geq 1] (S_{L_1}(t_{i,j}^{HR} - v_{i,j}) + S_{L_1}(t_{i,j}^{SR} - v_{i,j})), \quad (7)$$

in which,

$$S_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise,} \end{cases} \quad (8)$$

where (x, y, w, h) denotes the center, width, and height of a face region respectively, $v_i = (v_{i,x}, v_{i,y}, v_{i,w}, v_{i,h})$ denotes a tuple of ground truth bounding-box regression target, and $t_i = (t_{i,x}, t_{i,y}, t_{i,w}, t_{i,h})$ denotes a predicted regression tuple. t_i^{HR} and t_i^{SR} denote the predicted tuple for the i -th real high-resolution and generated super-resolved images, respectively. The bracket

indicator function $[u_i \geq 1]$ equals 1 when $u_i \geq 1$ and 0 otherwise. For more detailed descriptions of the regression loss, please refer to [52].

Similar to the classification loss, the introduced regression loss also plays two roles. First, it promotes the discriminator network to regress locations of the face candidate regions cropped from the baseline detector. Second, it drives the generator network to produce high quality super-resolved images with fine details for easier classification and better localization.

Objective function. Based on the above analysis, we incorporate the adversarial loss Eq(5), classification loss Eq(6) and regression loss Eq(7) into the pixel-wise MSE loss Eq(4) to train our proposed networks. The GAN-based network can be trained by the objective function Eq(9):

$$\begin{aligned} \max_{\theta} \min_w \frac{1}{N} \sum_{i=1}^N & \alpha (\log(1 - D_{\theta}(G_w(I_i^{LR}))) + \log D_{\theta}(I_i^{HR})) + \\ & \frac{1}{N} \sum_{i=1}^N \beta (\log(y_i - D_{cls}(G_w(I_i^{LR}))) + \log(y_i - D_{cls}(I_i^{HR}))) + \\ & \frac{1}{N} \sum_{i=1}^N \gamma \sum_{j \in \{x,y,w,h\}} [u_i \geq 1] (S_{L_1}(t_{i,j}^{HR} - v_{i,j}) + S_{L_1}(t_{i,j}^{SR} - v_{i,j})) + \\ & \frac{1}{N} \sum_{i=1}^N (\|G_{1_{w_1}}(I_i^{LR}) - I_i^{HR}\|^2 + \|G_{2_{w_2}}(G_{1_{w_1}}(I_i^{LR})) - I_i^{HR}\|^2), \end{aligned} \quad (9)$$

where α , β , and γ are trade-off weights, which are set by cross-validation experiments.

For better gradient behavior, we optimize the objective function in an alternative way as in [39, 40, 47] and modify the loss function of generator G and the discriminator D as:

$$\begin{aligned} \min_w \frac{1}{N} \sum_{i=1}^N & \alpha \log(1 - D_{\theta}(G_w(I_i^{LR}))) + \\ & \frac{1}{N} \sum_{i=1}^N \beta \log(y_i - D_{cls}(G_w(I_i^{LR}))) + \\ & \frac{1}{N} \sum_{i=1}^N \gamma \sum_{j \in \{x,y,w,h\}} [u_i \geq 1] (S_{L_1}(t_{i,j}^{HR} - v_{i,j}) + S_{L_1}(t_{i,j}^{SR} - v_{i,j})) + \\ & \frac{1}{N} \sum_{i=1}^N (\|G_{1_{w_1}}(I_i^{LR}) - I_i^{HR}\|^2 + \|G_{2_{w_2}}(G_{1_{w_1}}(I_i^{LR})) - I_i^{HR}\|^2), \end{aligned} \quad (10)$$

and

$$\begin{aligned}
\min_{\theta} & \frac{1}{N} \sum_{i=1}^N -\alpha((\log(1 - D_{\theta}(G_w(I_i^{LR}))) + \log D_{\theta}(I_i^{HR})) + \\
& \frac{1}{N} \sum_{i=1}^N \beta(\log(y_i - D_{cls}(G_w(I_i^{LR}))) + \log(y_i - D_{cls}(I_i^{HR}))) + \\
& \frac{1}{N} \sum_{i=1}^N \gamma \sum_{j \in \{x,y,w,h\}} [u_i \geq 1](S_{L_1}(t_{i,j}^{HR} - v_{i,j}) + S_{L_1}(t_{i,j}^{SR} - v_{i,j})).
\end{aligned} \tag{11}$$

The loss function of generator G in Eq(10) consists of adversarial loss Eq(5), MSE loss Eq(4), and classification loss Eq(6) & regression loss Eq(7), which enforce the reconstructed images to be similar to real natural high-resolution images on high-frequency details, pixel, and semantic level respectively. Different from the previous GANs, we add the classification and regression losses of generated super-resolved images to the objective loss of the generative network. By introducing these two losses to the generative network, the recovered images from generator are more realistic than the results optimized only by the adversarial loss and MSE loss. Further ablation analysis on the influence of each loss function is presented in subsection 4.3.

In the loss function of discriminator D in Eq(11), we introduce the classification loss (*i.e.* classification branch) to classify whether the high-resolution images are faces or non-faces, which is parallel to the basic formulation of GAN [35] to distinguish whether the high-resolution images are fake or real. For small faces detection, as we all know, a few pixel drift may result in predicted bounding-boxes fail to fulfill the evaluation criteria. Therefore, we further introduce the regression loss (*i.e.* regression branch) to the discriminator network.

3.5. Exploring Contextual Information

In the above subsections, we have shown an effective face detector based on GAN. However, some too small (*e.g.* 10×10 pixels) and too blurry faces may still be lost or some false positive results (*i.e.* background regions) may be predicted as a face. To find these hard faces which are not detected by the above GAN-based face detector and to alleviate those false positive results, we introduce the contextual information to train the proposed GAN-based networks, which is inspired from the intuition of human vision system. Specifically, when humans are finding a face, especially for the hard face (*e.g.* small, blurry, occluded, *etc.*), they always try to search not only the facial unit, but also some contextual parts like the hair, neck, even body. Moreover, people also use these contextual information to reject some false positive results (*i.e.* the backgrounds

Figure 4: Illustration of exploring contextual information. (a) is a training image with many tiny faces. (b) is a face region with different scale contextual information cropped from (a). In sub-image (c), “X1.0” means the face region itself generated by our baseline face detector [9] and no context is included here, “X1.5” denotes the face region with contextual information which is obtained by enlarging the crop window by a factor of 1.5, and “X2.0”, “X3.0”, “X4.0” are obtained by enlarging the crop window by the factor of 2.0, 3.0, 4.0, respectively. For visualization, we would like to note that we choose a medium face to show the method of exploring contextual information instead of a hard face. Best viewed on the computer, in color and zoomed in.

look like a face region). Based on this intuition, we propose a context-assisted hard face detection method that includes the contextual information when preparing the training data.

The training data, including the positive data (faces) and negative data (non-faces), of our proposed GAN-based face detector is cropped by our baseline MB-FCN detector, and no contextual information is included in the face region as shown in Figure 4 (c) “X1.0”, which leads to missing some hard faces. In order to utilize the contextual information around the face region, we enlarge the crop window by a factor $m = 1.5, 2.0, 3.0, 4.0$, see Figure 4 (c). With the help of contextual information, the hard face detection is getting easier, *i.e.* some missing faces can be found, and the false positive results from similar background regions can be alleviated effectively. Thus, the performance of detecting tiny faces in the wild will improve dramatically.

In the testing process, we choose the best factor m from 1.5, 2.0, 3.0, 4.0 by validation experiments (see ablation studies in subsection 4.3) to produce RoIs, and then use the trained discriminator model to produce final detection results, so the inference process is efficient.

4. Experiments

In this section, we experimentally validate our proposed method on two public face detection benchmarks (*i.e.* WIDER FACE [25] and FDDB [24]). First, we conduct an ablation experiment to prove the effectiveness of GAN. Then, we give a detailed analysis on the importance of each loss in the generator and discriminator networks, the influence of contextual information, and how sensitivity of our proposed method to the baseline detector. Finally, our proposed face detector is evaluated on both of these two public benchmarks while comparing the performance against other state-of-the-art approaches.

4.1. Training and Validation Datasets

We use a recently released large-scale face detection benchmark, WIDER FACE dataset [25], to train our proposed networks. It contains 32,203 images, which are selected from the publicly available WIDER dataset. 40%/10%/50% of the data is randomly selected for training, validation, and testing, respectively. Images in WIDER FACE are categorized into 61 social event classes, which have much more diversities and are closer to the real-world scenario. Therefore, we use this dataset for training and validating the proposed generator and discriminator networks.

The WIDER FACE dataset is divided into three subsets, Easy, Medium, and Hard, based on the heights of the ground truth faces. The Easy/Medium/Hard subset contains faces with heights larger than 50/30/10 pixels respectively. Compared to the Medium subset, the Hard one contains many faces with a height between 10–30 pixels. As expected, it is quite challenging to achieve good detection performance on the Hard subset.

4.2. Implementation Details

In the generator network, we set trade-off weights $\alpha = 0.001$ and $\beta = \gamma = 0.01$. During training, we use the Adam optimizer with momentum term $\beta_1 = 0.9$. The generator network is trained from scratch and the weights in each layer are initialized with a zero-mean Gaussian distribution with standard deviation 0.02, and biases are initialized with 0. To avoid undesirable local optima, we first train an MSE-based SR network to initialize the generator network. For the discriminator network, we employ the ResNet50 [15] pre-trained on ImageNet as our backbone network and we replace all the fc layers with three parallel fc layers. The fc layers are initialized by a zero-mean Gaussian distribution with standard deviation 0.1, and biases are initialized with 0.

Our baseline MB-FCN detector is based on Resnet50 network [15], which is pre-trained on ImageNet. All hyper-parameters of the MB-FCN detector are the same as [9]. For training our generator and discriminator networks, we crop 1,075,968 face samples and 1,626,328 non-face samples from WIDER FACE [25] training set with our baseline detector. The corresponding low-resolution images are generated by down-sampling the high-resolution images using the bicubic interpolation method with a factor 4. During testing, 600 regions of interest (ROIs) are cropped by our baseline detector, and these ROIs are fed to our GAN-based networks to give out the final detection performance.

The GAN-based networks are trained with first 6 epochs at a learning rate of 10^{-2} and another 4 epochs at a lower learning rate of 10^{-3} . We alternately update the generator and discriminator network, which is equivalent to $k = 1$ as in [35]. Our implementation is based on tensorflow, and all the experiments are done on an EVIDIA TITAN X GPU.

4.3. Ablation Studies

We first compare our proposed method with the baseline detector to prove the effectiveness of the GAN architecture. Then, we perform an ablation study by removing the refinement network to testify the positive influence of refinement network. Moreover, we validate the function of the regression branch in the discriminator network by comparing the detection performance with/without this branch. Furthermore, to validate the contribution of each loss, including adversarial loss, classification loss, and regression loss in the loss function of generator network, we also conduct ablation studies by cumulatively adding each of them to the pixel-wise loss. Finally, we also design the experiments to prove the effectiveness of the contextual information.

Influence of the GAN. Table 2 (the 1st and the 5th row) shows the detection performance (AP) of the baseline detector and our method on WIDER FACE validation set. Our baseline detector is a multi-branch RPN face detector with skip connection of feature maps. From Table 2 we observe that the performance of our proposed detector outperforms the baseline detector by a large margin (1.5% in AP) on the Hard subset. The reason is that the baseline detector performs down-sampling operations (*i.e.* convolution with stride 2) on the small faces. The small faces themselves contain limited information, and the majority of detailed information will be lost after several convolutional operations. For example, the input is a 16×16 pixels face, and the result is 1×1 on the C4 feature map and nothing is reserved on the C5 feature map. Based

Table 2: Performance of the baseline model trained with and without GAN, refinement network, regression branch, adversarial loss, classification loss, and regression loss on the WIDER FACE validation set. “adv loss” denotes adversarial loss Eq(5), “cls loss” represents classification loss Eq(6), “MES loss” means pixel-wise loss Eq(4), “reg+ branch” denotes the regression branch in the discriminator network, and “reg loss” denotes the regression loss Eq(7) in the discriminator network, which is back-propagated to the loss function of generator network to drives the generator to recover some fine details for better localization.

Method	Easy	Medium	Hard
Baseline[9]	0.932	0.922	0.858
w/o Refinement Network	0.940	0.929	0.863
w/o adv loss	0.935	0.925	0.867
w/o cls loss	0.936	0.927	0.865
Ours(Baseline + MES loss + adv loss + cls loss)	0.944	0.933	0.873
w/o reg+ branch	0.946	0.936	0.876
w/o reg loss	0.949	0.940	0.880
Ours*(Ours + (reg+ branch) + reg loss)	0.951	0.943	0.882

on those limited features, it is normal to get the poor detection performance. In contrast, our method first learns a super-resolution image and then refines it, which solves the problem that the original small blurry faces lack detailed information and blurry simultaneously. Based on the super-resolved images with fine details, the boost of detection performance is inevitable.

Influence of the refinement network. From Table 2 (the 2nd and 5th row), we see that the AP performance increases by 1% on the Hard subset by adding the refinement sub-network to the generator network. Interestingly, the performances of Easy and Medium subsets also have an improvement (0.4%). We visualize the reconstructed faces from the generator network and find that our refinement network can reduce the influence of illumination and blur as shown in Figure 5. In some cases, the baseline detector fails to detect the faces if those faces are heavily blurred or illuminated. However, our method reduces influence of such attributions and can find these faces successfully. Here, we would like to note that our framework is not specific and any off-the-shelf face detectors can be used as our baseline detector.

Influence of the regression branch. As shown in Figure 1, the detection of some hard faces are failed without the regression branch. Actually, imperfect localization is one of the main

Figure 5: Some examples of clear faces generated by our generator network from the blurry ones. The top row shows the small faces influenced by the factors of blur and illumination, and the bottom row shows the clearer faces generated by our method. The low-resolution images in the top row are re-sized for visualization.

reasons for detection errors, especially for the small cases where small shifts in their bounding boxes will lead to a failed detection when the strict evaluation metrics is used. To overcome this issue, we further introduce the regression branch in the discriminator to refine the predicted bounding boxes. From Table 2 (6th and 8th row), we can see that the performance on the Hard subset improves by 0.6% when the regression branch is added, thus demonstrating its effectiveness in the proposed pipeline.

Influence of the adversarial loss. From Table 2 (the 3rd and 5th row), we see that the AP performance drops by about 1% without the adversarial loss. The reason is that the generated images derived by pixel-wise loss and classification loss are over smooth. Upon closely inspecting the generated images, we find that the fine details around eyes are of low quality. Since these details are not important features for the discriminator to distinguish fake/real images, and the generator can still fool the discriminator when making mistakes in this region. To encourage the generator to restore the high-quality images, we include the adversarial loss in the loss function of generator network.

Influence of the classification loss. From Table 2 (the 4th and 5th row), we see that the AP performance increases by about 1% by adding the classification loss. This is because the classification loss promotes the generator to recover fine details for easier classification. We find that the generated faces have clearer contour when adding the classification loss. We think the contour information may be the most important evidence for the discriminator to classify faces/non-faces when faces are too small and heavily blurred. So we add the classification loss to our generator loss function.

Influence of the regression loss. To validate the effectiveness of introducing the regression

Figure 6: The AP performance with different magnification factor of the crop window when preparing the training data on WIDER FACE *validation* set.

loss to the generator network, we also conduct an ablation experiment with and without this loss in the objective loss function of the generator network. As shown in Table 2 (the 7th and 8th row), the performance increases by 0.2% and 0.3% absolutely on the Hard and Medium subsets respectively when using the regression loss to train the generator network. Similar to the classification loss, the regression loss drives the generator to recover some fine details for better localization. The increased AP demonstrates the necessity of introducing the regression loss to the generator loss function, therefore, we further add the regression loss to the generator loss function in our proposed method.

Influence of the contextual information. With the rich contextual information, it is easy to find the tiny faces and alleviate the false positive results. Therefore, it is expected that the detection performance would increase obviously. Figure 6 shows the comparison of AP performance of different magnification factor of the crop window on WIDER FACE validation set. It can be observed that the AP performance of original cropped face regions without contextual information (*i.e.* X1.0) is 95.1%/94.3%/88.2% on Easy/Medium/Hard subset, respectively. By enlarging the crop window by a factor 1.5, the detection performance increases slightly. This demonstrates the contextual information has a positive influence on face detection in the wild. By continuing to enlarge the crop window to utilize the contextual information around face regions, the AP performance shows a continuous growth trend. However, the performance will decrease

Table 3: The performance (AP) in different magnification factors of the crop window on the WIDER FACE validation set, where “Xm” represents the different magnification factor of the crop window.

Method	Easy	Medium	Hard
X1.0	0.951	0.943	0.882
X1.5	0.953	0.944	0.886
X2.0	0.955	0.946	0.889
X3.0	0.960	0.949	0.872
X4.0	0.936	0.925	0.814

dramatically when the crop window becomes too large (*e.g.*, X3.0, X4.0). By visualizing the prepared training data, as shown in Figure 4 (c), we find more than one face or extra backgrounds are included in the cropped regions, which looks more like the background than a face. We think the reason of AP performance decreases is that some noises are brought into the cropped regions.

To better understand the contribution of the contextual information quantitatively, Table 3 shows the comparison of each magnification factor m ($m=1.0, 1.5, 2.0, 3.0, 4.0$) of the crop window to generate the training data for training the GAN-based networks on WIDER FACE validation set. We can see that the detection performance indeed increases, and the AP performance improves to 95.5%/94.6%/88.9% at X2.0, to 96.0%/94.9%/87.2% at X3.0 on Easy/Medium/Hard subset, respectively. For the Easy and Medium subset, we get the best performance when the magnification factor of the crop window is set to 3.0 (*i.e.* X3.0). In terms of the Hard subset, however, it reaches the peak when the magnification factor equals 2.0 (*i.e.* X2.0). The reason is that the generated training data by the crop window X3.0 usually includes more than one face or extensive background for Hard subset, but the contextual information cropped by X3.0 is suitable for Easy and Medium subset. The purpose of our paper is to detect tiny faces, so we set the magnification factor to 2.0 (*i.e.* X2.0) for all the experiments.

How sensitive is the proposed method to the baseline detector? To validate the robustness of the proposed face detector, we use a weaker baseline detector to replace the current baseline detector [9], and test the detection performance on WIDER FACE validation set. More concretely, we employ the model in [47] as a weaker WSD, whose performance is 91.9%/90.8%/82.3% on the Easy/Medium/Hard subset respectively, decreasing by more than 3%

Figure 7: On the WIDER FACE validation set, we compare our method with several state-of-the-art methods: MSCNN[25], MTTCNN[53], CMS-RCNN[11], HR[47], SSH[48], SFD[54], Face R-FCN[55], and Zhu *et. al* [56]. The average precision (AP) is reported in the legend. Best viewed in color and zoomed in.

absolutely compared with our baseline detector. The weaker baseline detector is used to prepare the training data, and to generate the candidate RoIs for testing. As a result, our proposed framework can produce a comparable performance 95.3%/94.5%/88.8% on the Easy/Medium/Hard subset respectively even though we replace the current superior baseline detector with the weaker one, which verifies that our proposed method is not sensitive to the baseline face detector. We would like to note that the only thing needs to be ensured is that the baseline detector needs a high recall rate. To be our best knowledge, any off-the-shelf face detectors with high recall can be used as our baseline detector here.

4.4. Comparison with the State-of-the-Art Methods

We compare our proposed method with previous state-of-the-art methods on two public face detection benchmarks (*i.e.* WIDER FACE [25] and FDDB [24]).

Evaluation on WIDER FACE validation set. We compare our method with the previous state-of-the-art face detectors [25, 53, 11, 47, 48, 54, 55, 56]. Figure 7 shows the performance on WIDER FACE validation set. From Figure 7, we see that our method achieves the highest performance (*i.e.* 88.9%) on the Hard subset, outperforming the state-of-the-art face detector by more than 1% AP absolutely. Compared to these CNN-based methods, the boost of our performance mainly comes from four contributions: (1) our up-sampling sub-network in the generator learns a super-resolution image, which reduces too much information loss caused by down-sampling while implementing convolution operations on those small faces; (2) the refinement sub-network in the generator learns finer details and reconstructs clearer images. Based on the clear super-

Figure 8: On the WIDER FACE test set, we compare our proposed method with several previous state-of-the-art methods: MSCNN[25], MTTCNN[53], CMS-RCNN[11], HR[47], SSH[48], SFD[54], Face R-FCN[55], and Zhu *et al* [56]. The average precision (AP) is reported in the legend. Best viewed in color and zoomed in.

resolution images, it is easier for the discriminator to classify faces or non-faces than depending on low-resolution blurry images; (3) the classification loss Eq(6) and regression loss Eq(7) promote the generator to learn a clearer face for easier classification and better localization; (4) the regression branch and contextual information are further introduced to detect the hard faces. Furthermore, we also get the highest performance (94.6%/95.5%) on the Easy/Medium subset, outperforming the state-of-the-art face detector by 0.6% and 1.1% respectively. This is because some big faces are heavily influenced by illumination and blur, as shown in Figure 5. As a result, these CNN-based methods fail to detect these faces. However, our method alleviates the influence of these attributions and finds these faces successfully.

After using the contextual information, the AP performance further reaches 94.9%/94.2%/88.9% on Easy/Medium/Hard subset, respectively, outperforming the state-of-the-art face detector by a large margin. Especially on Hard subset, the absolute AP performance increases by 3.6% when comparing with the previous best face detector, which certifies the effectiveness of contextual information in detecting the tiny faces in the wild.

Evaluation on WIDER FACE test set. Figure 8 shows the precision-recall curves on WIDER FACE test set. Similarly, our proposed method achieves the best result on the Easy/Medium/Hard subset, *i.e.* 0.951 (Easy), 0.942 (Medium) and 0.883 (Hard). Considering the Hard subset contains lots of occluded faces, small faces and blurry faces, our method outperforms the previous state-of-the-art method by a large margin, more than 1% AP absolutely, which further validates the effectiveness of our proposed method in handling small blurry faces

Figure 9: On the FDDB dataset, we compare our method against some previous state-of-the-art methods. The precision rate with 500 false positives is reported. Best viewed in color and zoomed in.

in the wild.

Evaluation on FDDB dataset. We follow the standard metrics (*i.e.* precision at specific false positive rates) of the FDDB [24] and use this metric to compare with other methods. There are many unlabeled faces in FDDB, making precision not accurate at small false positive rates. Hence, we report the precision rate at 500 false positives. Our face detector achieves a superior performance (0.973) over all other state-of-the-art face detectors except SFD [54] detector, as shown in Figure 9. We would like to note that the performance of SFD [54] is achieved after manually adding 238 unlabeled faces on the test set. However, we test our model on the original labeled test set. Under such an unfair condition, our method still gets the comparable performance, which further proves the effectiveness of our method.

Since most of the faces in FDDB database are large and few faces are blurry or obscured, there is no significant improvement in AP performance by including contextual information and the regression branch, so we do not show the detection results of adding contextual information and the regression branch on FDDB dataset in our paper.

4.5. Qualitative Results

In Figure 10, we show some detection results generated by our proposed method. It can be found that our face detector successfully finds almost all the faces, even though some faces are very small and heavily blurred. However, Figure 10 also shows some failure cases, *e.g.* some false positive results. By carefully inspecting the detection results, we find that these failure cases

Figure 10: Qualitative detection results of our proposed method. Green bounding boxes are ground truth annotations and red bounding boxes are the results predicted by our proposed method. Best seen on the computer, in color and zoomed in.

are some objects that are very similar to a small blurry face. These results indicate that more progress is needed to further improve the small face detection performance.

5. Conclusion and Future Work

In this paper, we propose a new method by using GAN to detect small faces in the wild. In the generator network, we design a novel network to directly generate a clear super-resolution image from a small blurry one, and our up-sampling sub-network and refinement sub-network are trained in an end-to-end way. Moreover, contextual information is further utilized to find the hard faces and alleviate the false positive results. Extensive experiments on WIDER FACE and Fddb demonstrate the substantial improvements of our method in the Hard subset, as well as in the Easy/Medium subset, when compared to previous state-of-the-art face detectors. We would like to note that we are the first to introduce the super-resolution technique to solve the problem of detecting small faces in the wild. Base on the proposed method, we can detect small faces in unconstrained scenarios (*e.g.* 12×12 pixels faces in the wild) rather than in ideal scenes (*i.e.* large faces which are located in the center of an image). However, we still cannot solve some

extremely hard faces, *e.g.* some backgrounds are very similar to a small blurry face. We plan to use structure information of faces (*i.e.* landmarks) to detect those failure cases in the future.

Furthermore, combining GAN and super-resolution technique is not limited to detecting tiny faces in the wild, and it can be extended to detect any general small objects in the unconstrained conditions. In the future, we plan to use our proposed GAN-based network to solve the problem of detecting general small objects in unconstrained scenarios, *e.g.* detecting 80 categories general objects on the COCO [57] dataset.

Acknowledgment

The majority of this work was done when Yongqiang Zhang worked at King Abdullah University of Science and Technology (KAUST) as a visiting PhD student, and continued at Harbin Institute of Technology (HIT). This work was supported by Natural Science Foundation of China, Grant No. 61603372.

References

References

- [1] Y. Li, C. Cai, G. Qiu, K.-M. Lam, Face hallucination based on sparse local-pixel structure, *Pattern Recognition* 47 (3) (2014) 1261–1270.
- [2] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, Y. Ma, Robust face recognition via sparse representation, *IEEE transactions on pattern analysis and machine intelligence* 31 (2) (2009) 210–227.
- [3] J. Ma, J. Zhao, Y. Ma, J. Tian, Non-rigid visible and infrared face registration via regularized gaussian fields criterion, *Pattern Recognition* 48 (3) (2015) 772–784.
- [4] W. Ou, X. You, D. Tao, P. Zhang, Y. Tang, Z. Zhu, Robust face recognition via occlusion dictionary learning, *Pattern Recognition* 47 (4) (2014) 1559–1572.
- [5] Y. Gao, Y. Qi, Robust visual similarity retrieval in single model face databases, *Pattern Recognition* 38 (7) (2005) 1009–1020.
- [6] Z. Wu, Q. Ke, J. Sun, H.-Y. Shum, Scalable face image retrieval with identity-based quantization and multireference reranking, *IEEE transactions on pattern analysis and machine intelligence* 33 (10) (2011) 1991–2001.

- [7] S. Zafeiriou, C. Zhang, Z. Zhang, A survey on face detection in the wild: past, present and future, *Computer Vision and Image Understanding* 138 (2015) 1–24.
- [8] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, M.-H. Yang, Learning to super-resolve blurry face and text images, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 251–260.
- [9] Y. Bai, B. Ghanem, Multi-branch fully convolutional network for face detection, *arXiv preprint arXiv:1707.06330*.
- [10] H. Jiang, E. Learned-Miller, Face detection with the faster r-cnn, in: *FG 2017*, 2017, pp. 650–657.
- [11] C. Zhu, Y. Zheng, K. Luu, M. Savvides, Cms-rcnn: contextual multi-scale region-based cnn for unconstrained face detection, in: *Deep Learning for Biometrics*, Springer, 2017, pp. 57–79.
- [12] Y. Bai, Y. Zhang, M. Ding, B. Ghanem, Finding tiny faces in the wild with generative adversarial network, *CVPR*. IEEE.
- [13] X. Tang, D. K. Du, Z. He, J. Liu, Pyramidbox: A context-assisted single shot face detector, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 797–813.
- [14] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556*.
- [15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [16] J. Yan, Z. Lei, L. Wen, S. Z. Li, The fastest deformable part model for object detection, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2497–2504.
- [17] J. Yan, X. Zhang, Z. Lei, S. Z. Li, Real-time high performance deformable model for face detection in the wild, in: *2013 International Conference on Biometrics (ICB)*, IEEE, 2013, pp. 1–6.
- [18] Y. Zhang, M. Ding, W. Fu, Y. Li, Reading recognition of pointer meter based on pattern recognition and dynamic three-points on a line, in: *ICMV 2016*, Vol. 10341, International Society for Optics and Photonics, 2017, p. 103410K.

- [19] P. Koppen, Z.-H. Feng, J. Kittler, M. Awais, W. Christmas, X.-J. Wu, H.-F. Yin, Gaussian mixture 3d morphable face model, *Pattern Recognition* 74 (2018) 617–628.
- [20] L. Zhang, L. Lin, X. Liang, K. He, Is faster r-cnn doing well for pedestrian detection?, in: *European conference on computer vision*, Springer, 2016, pp. 443–457.
- [21] Z. Cai, Q. Fan, R. S. Feris, N. Vasconcelos, A unified multi-scale deep convolutional neural network for fast object detection, in: *European conference on computer vision*, Springer, 2016, pp. 354–370.
- [22] Y. Zhang, Y. Bai, M. Ding, Y. Li, B. Ghanem, W2f: A weakly-supervised to fully-supervised framework for object detection, in: *CVPR. IEEE*, 2018, pp. 928–936.
- [23] Y. Zhang, Y. Bai, M. Ding, Y. Li, B. Ghanem, Weakly-supervised object detection via mining pseudo ground truth bounding-boxes, *Pattern Recognition* 84 (2018) 68–81.
- [24] V. Jain, E. Learned-miller, Fddb: A benchmark for face detection in unconstrained settings, *Tech. rep.*, a (2010).
- [25] S. Yang, P. Luo, C.-C. Loy, X. Tang, Wider face: A face detection benchmark, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5525–5533.
- [26] J. Shi, X. Liu, Y. Zong, C. Qi, G. Zhao, Hallucinating face image by regularization models in high-resolution feature space, *IEEE Transactions on Image Processing* 27 (6) (2018) 2980–2995.
- [27] X. Yu, B. Fernando, R. Hartley, F. Porikli, Super-resolving very low-resolution face images with supplementary attributes, in: *CVPR. IEEE*, 2018, pp. 908–917.
- [28] E. Ribeiro, A. Uhl, F. Alonso-Fernandez, Iris super-resolution using cnns: is photo-realism important to iris recognition?, *IET Biometrics* 8 (1) (2018) 69–78.
- [29] Q. Shan, J. Jia, A. Agarwala, High-quality motion deblurring from a single image, *Acm transactions on graphics (tog)* 27 (3) (2008) 73.
- [30] X. Li, Y.-q. Zhang, X.-t. Yang, Study of a self-adaptive kalman filter method in ngmimu/gps integrated navigation scheme, in: *Proceedings of the 6th International Asia Conference on Industrial Engineering and Management Innovation*, Springer, 2016, pp. 225–235.
- [31] F. Zhao, J. Feng, J. Zhao, W. Yang, S. Yan, Robust lstm-autoencoders for face de-occlusion in the wild, *IEEE Transactions on Image Processing* 27 (2) (2018) 778–790.

- [32] J. Sun, W. Cao, Z. Xu, J. Ponce, Learning a convolutional neural network for non-uniform motion blur removal, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 769–777.
- [33] Y. Zhang, S. Peng, D. Mingli, N. Xue, L. Yongqiang, L. Xian, Design and implementation of a test platform for the preloaded assembly of harmonic reducer, in: 2015 12th IEEE International Conference on Electronic Measurement & Instruments (ICEMI), Vol. 1, IEEE, 2015, pp. 522–526.
- [34] A. Chakrabarti, A neural approach to blind motion deblurring, in: European conference on computer vision, Springer, 2016, pp. 221–235.
- [35] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems 27, Curran Associates, Inc., 2014, pp. 2672–2680.
- [36] E. L. Denton, S. Chintala, a. szlam, R. Fergus, Deep generative image models using a laplacian pyramid of adversarial networks, in: Advances in Neural Information Processing Systems 28, Curran Associates, Inc., 2015, pp. 1486–1494.
- [37] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, A. A. Efros, Generative visual manipulation on the natural image manifold, in: European Conference on Computer Vision, Springer, 2016, pp. 597–613.
- [38] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, Y. LeCun, Disentangling factors of variation in deep representation using adversarial training, in: Advances in Neural Information Processing Systems 29, Curran Associates, Inc., 2016, pp. 5040–5048.
- [39] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4681–4690.
- [40] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134.
- [41] M. Mueller, N. Smith, B. Ghanem, Context-aware correlation filter tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1396–1404.

- [42] D. Park, D. Ramanan, C. Fowlkes, Multiresolution models for object detection, in: ECCV, Springer, 2010, pp. 241–254.
- [43] S. Bell, C. Lawrence Zitnick, K. Bala, R. Girshick, Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks, in: CVPR. IEEE, 2016, pp. 2874–2883.
- [44] X. Zeng, W. Ouyang, J. Yan, H. Li, T. Xiao, K. Wang, Y. Liu, Y. Zhou, B. Yang, Z. Wang, et al., Crafting gbd-net for object detection, IEEE transactions on pattern analysis and machine intelligence 40 (9) (2018) 2109–2123.
- [45] S. Gidaris, N. Komodakis, Object detection via a multi-region and semantic segmentation-aware cnn model, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1134–1142.
- [46] H. Hu, J. Gu, Z. Zhang, J. Dai, Y. Wei, Relation networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3588–3597.
- [47] P. Hu, D. Ramanan, Finding tiny faces, in: CVPR, 2017 IEEE Conference on, IEEE, 2017, pp. 1522–1530.
- [48] M. Najibi, P. Samangouei, R. Chellappa, L. S. Davis, Ssh: Single stage headless face detector., in: ICCV, 2017, pp. 4885–4894.
- [49] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 25, Curran Associates, Inc., 2012, pp. 1097–1105.
- [50] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv preprint arXiv:1511.06434.
- [51] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: F. Bach, D. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, Vol. 37 of Proceedings of Machine Learning Research, PMLR, Lille, France, 2015, pp. 448–456.
- [52] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

- [53] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, *IEEE Signal Processing Letters* 23 (10) (2016) 1499–1503.
- [54] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, S. Z. Li, S³fd: Single shot scale-invariant face detector, in: *Computer Vision (ICCV), 2017 IEEE International Conference on*, IEEE, 2017, pp. 192–201.
- [55] Y. Wang, X. Ji, Z. Zhou, H. Wang, Z. Li, Detecting faces using region-based fully convolutional networks, *arXiv preprint arXiv:1709.05256*.
- [56] C. Zhu, R. Tao, K. Luu, M. Savvides, Seeing small faces from robust anchors perspective, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5127–5136.
- [57] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: *ECCV*, Springer, 2014, pp. 740–755.

Author biographies

Yongqiang Zhang received the MS degree in instrument science and technology from Harbin Institute of Technology, Harbin, China, in 2013. He is a Ph.D. student from the Harbin Institute of Technology (HIT). He worked as a visiting student at King Abdullah University of Science and Technology (KAUST) from 2017 to 2018. His research areas are computer vision, pattern recognition, and deep learning. His research interests mainly include face detection, object detection, activity detection, image and video understanding in the real-world.

Mingli Ding received the BS, MS and PhD degrees in instrument science and technology from Harbin Institute of Technology (HIT), Harbin, China, in 1996, 1997 and 2001, respectively. Currently, he is a professor at Harbin Institute of Technology. Prof. Ding's research interests are intelligence tests and information processing, automation test technology, and computer vision. He has published over 40 papers in peer-reviewed journals and conferences.

Yancheng Bai is an associate professor in Pattern Recognition and Intelligent System at Institute of Software, Chinese Academy of Sciences. He held a Postdoctoral position at King Abdullah University of Science and Technology (KAUST) from 2016 to 2018. He received the PhD degree from Institute of Automation, Chinese Academy of Sciences in 2014. His research interests include computer vision, pattern recognition, artificial intelligence, and machine learning.

Bernard Ghanem is currently an assistant professor in the CEMSE division and a member of the Visual Computing Center at KAUST. Before that, he was a senior research scientist at the University of Illinois Urbana-Champaign (UIUC) in Singapore, where he still holds an adjunct position. He heads projects that develop algorithms in computer vision, machine learning, and optimization geared towards real-world applications, including semantic video analysis in sports and automated surveillance, content-based image retrieval, large-scale activity recognition, and 2D/3D scene understanding. He received his Bachelor's degree in Computer and Communications Engineering from the American University of Beirut (AUB) in 2005 and his MS/PhD in Electrical and Computer Engineering from UIUC in 2010. His work has received several awards and honors, including the Henderson Graduate Award from UIUC, two consecutive CSE fellowship awards from UIUC, a Best Paper Award (CVPRW 2013), a two-year KAUST Seed Fund, and a Google Faculty Research Award in 2015. He has co-authored more than 50 peer reviewed conference and journal papers in his field, as well as, 4 patents. He is also a co-founder of AutoScout Inc. that provides automated solutions for sports video analytics.